Forecasting with a Noncausal VAR Model

Nyberg, Henri

2014-08

Nyberg , H & Saikkonen , P 2014 , ' Forecasting with a Noncausal VAR Model ' , Computational Statistics & Data Analysis , vol. 76 , pp. 536-555 . https://doi.org/10.1016/j.csda.2013.10.014

http://hdl.handle.net/10138/223796
https://doi.org/10.1016/j.csda.2013.10.014

acceptedVersion

Forecasting with a Noncausal VAR Model

Henri Nyberg[a,*], Pentti Saikkonen[b,c]

[a] *Department of Political and Economic Studies, Economics, University of Helsinki, Arkadiankatu 7 (P.O.Box 17), 00014 University of Helsinki, Finland*
[b] *Department of Mathematics and Statistics, University of Helsinki, Finland*
[c] *Bank of Finland*

Abstract

Simulation-based forecasting methods for a non-Gaussian noncausal vector autoregressive (VAR) model are proposed. In noncausal autoregressions the assumption of non-Gaussianity is needed for reasons of identifiability. Unlike in conventional causal autoregressions the prediction problem in noncausal autoregressions is generally nonlinear, implying that its analytical solution is unfeasible and, therefore, simulation or numerical methods are required in computing forecasts. It turns out that different special cases of the model call for different simulation procedures. Monte Carlo simulations demonstrate that gains in forecasting accuracy are achieved by using the correct noncausal VAR model instead of its conventional causal counterpart. In an empirical application, a noncausal VAR model comprised of U.S. inflation and marginal cost turns out superior to the best-fitting conventional causal VAR model in forecasting inflation.

Keywords: Noncausal vector autoregression, forecasting, simulation, importance sampling, inflation.

1. Introduction

The conventional vector autoregressive (VAR) model has become a standard tool in various fields of applications. In economics and finance the VAR model is typically used in structural analysis to study the dynamics and interrelationships

*Corresponding author. Tel: +358-503182262, fax +358-09-19128736.
Email addresses: henri.nyberg@helsinki.fi (Henri Nyberg), pentti.saikkonen@helsinki.fi (Pentti Saikkonen)

between variables of interest. Another application of the VAR model is forecasting. For instance, economic decision makers, such as central banks and investors in financial markets, aim to forecast key macroeconomic and financial time series to assess the future state of the economy and investment opportunities.

The conventional causal VAR model has a moving average representation in terms of its present and past error terms. A characteristic feature of this model is that its error terms are not predictable by past values of the involved time series. In contrast, the moving average representation of the non-Gaussian noncausal VAR model recently considered by Davis and Song (2010) and Lanne and Saikkonen (2013) also involve future error terms that are predictable by past values of the considered time series. In addition to theoretical advancements these authors demonstrate the practical usefulness of the noncausal VAR model in economic and financial applications. As discussed by Lanne and Saikkonen (2013), an important economic application of the noncausal VAR model is checking the validity of widely used test procedures based on the causal VAR model in testing economic hypotheses, especially in models involving expectations.

As yet, the development of the noncausal VAR model is at its early stages and even the literature of univariate noncausal autoregressive models is scant (see Breidt et al. (1991), Rosenblatt (2000), Davis and Song (2010), Lanne and Saikkonen (2011, 2013) and the references therein). As demonstrated in this previous literature, noncausal autoregressions can be distinguished from their causal counterparts only when the data generation process is non-Gaussian. In noncausal autoregressions non-Gaussianity can therefore be seen as a necessary identification condition. The object of this paper is to devise forecasting techniques for the non-Gaussian noncausal VAR model of Lanne and Saikkonen (2013). In addition to computing forecasts these techniques are also needed in computing impulse response functions, and hence in conducting structural analysis within the noncausal VAR model. Thus, our contribution should widen the applicability of the noncausal VAR model in empirical research.

In the causal VAR model, forecasting is simple in that explicit formulas are

available. In the noncausal VAR model the situation is different because the prediction problem is, in general, nonlinear and, consequently, forecasts cannot be obtained without resorting to numerical methods. Further discussion on this point is provided by Lanne, Luoto, and Saikkonen (2012b) who develop a simulation-based forecasting method for the univariate noncausal AR model proposed by Lanne and Saikkonen (2011). It turns out that forecasts of the noncausal VAR model considered in this paper can be computed analogously only when a suitable condition on the structure of the model holds. One case where the required condition always holds is the purely noncausal VAR model whose moving average representation only involves present and future error terms. In general, the required condition states that a certain parameter matrix involving the autoregressive coefficients of the model is nonsingular. Due to estimation errors this nonsingularity always holds in practice but, to avoid potential problems with nearly singular cases, we develop a forecasting technique which does not depend on the structure of the model. To achieve this robustness, more demanding computations based on importance sampling are needed. A somewhat similar technique has recently been used by Breidt and Hsu (2005) in forecasting non-Gaussian and potentially noninvertible (univariate) moving average processes (for a general discussion of importance sampling, see, e.g., Geweke (1996)).

We examine the properties of our forecasting techniques by means of Monte Carlo simulations which also provide guidance for some user-chosen quantities needed in the application of these techniques. The simulations conducted demonstrate that our forecasting techniques perform well and that the correct noncausal VAR model outperforms its causal counterpart in forecast accuracy.

Although empirical experience of noncausal VAR models is still very limited, the findings of Lanne, Nyberg, and Saarinen (2012c) based on applying univariate autoregressions to a large economic data set suggest that noncausality is quite prevalent among economic time series (see also Lof (2013)). The related work of Lanne, Luoma, and Luoto (2012a) and Lanne et al. (2012b) complement these findings by demonstrating that the univariate noncausal AR model outperforms

its conventional causal counterpart in forecasting U.S. inflation. Our empirical application to inflation forecasting is partly motivated by the work of these previous authors. We consider a bivariate system consisting of inflation and the real marginal cost that has often been employed in monetary economics, especially in studies related to the New Keynesian Phillips Curve (see, e.g., Gali and Gertler (1999), Nason and Smith (2008), and the references therein). Our results are similar to those obtained by Lanne et al. (2012a, 2012b). We find that a noncausal VAR model provides the best in-sample fit and outperforms the best-fitting causal VAR model in out-of-sample forecasting.

The rest of the paper is structured as follows. Section 2 describes the noncausal VAR model of Lanne and Saikkonen (2013) and briefly discusses statistical inference. Section 3 develops the forecasting techniques of the paper, while Section 4 illustrates their performance by Monte Carlo simulations. Section 5 presents the empirical application. Section 6 concludes. Finally, some technical details are collected in four appendices.

2. Noncausal VAR model

In this section, we first describe the noncausal VAR model of Lanne and Saikkonen (2013) and then discuss briefly parameter estimation and statistical inference. Unless otherwise indicated, all vectors will be treated as column vectors and, for notational convenience, we shall write $x = (x_1, ..., x_n)$ for the (column) vector x where the components x_i may be either scalars or vectors (or both).

2.1. Model

Following Lanne and Saikkonen (2013) we consider the n-dimensional stochastic process y_t $(t = 0, \pm 1, \pm 2, ...)$ generated by

$$\Pi(B)\,\Phi\left(B^{-1}\right) y_t = \epsilon_t, \tag{1}$$

where ϵ_t $(n \times 1)$ is a sequence of independent, identically distributed random vectors with zero mean and finite positive definite covariance matrix, and $\Pi(B) = I_n - \Pi_1 B - \cdots - \Pi_r B^r$ and $\Phi(B^{-1}) = I_n - \Phi_1 B^{-1} - \cdots - \Phi_s B^{-s}$ are $n \times n$

matrix operators with B the usual backward shift operator, that is, $B^k y_t = y_{t-k}$ $(k = 0, \pm 1, ...)$. Moreover, the determinants of the matrix polynomials $\Pi(z)$ and $\Phi(z)$ $(z \in \mathbb{C})$ have their zeros outside the unit disc, so that

$$\det \Pi(z) \neq 0, \quad |z| \leq 1, \quad \text{and} \quad \det \Phi(z) \neq 0, \quad |z| \leq 1. \tag{2}$$

These conditions guarantee the validity of various moving average representations to be used in our subsequent developments.

If $\Phi_j \neq 0$ for some $j \in \{1, .., s\}$, equation (1) defines a noncausal vector autoregression referred to as purely noncausal when $\Pi_1 = \cdots = \Pi_r = 0$ (or $r = 0$). When $\Phi_1 = \cdots = \Phi_s = 0$ (or $s = 0$) the conventional causal model is obtained. Then the former condition in (2) guarantees the stationarity of the model. In the general set-up of model (1) the same is true for the process

$$u_t = \Phi\left(B^{-1}\right) y_t. \tag{3}$$

Specifically, there exists a $\delta_1 > 0$ such that $\Pi(z)^{-1}$ has a well defined power series representation $\Pi(z)^{-1} = \sum_{j=0}^{\infty} M_j z^j = M(z)$ for $|z| < 1 + \delta_1$. Consequently, the process u_t has the causal moving average representation

$$u_t = M(B)\,\epsilon_t = \sum_{j=0}^{\infty} M_j \epsilon_{t-j}, \tag{4}$$

where $M_0 = I_n$ and the coefficient matrices M_j decay to zero at a geometric rate as $j \to \infty$.

Write $\Pi(z)^{-1} = \det(\Pi(z))^{-1} \Xi(z) = M(z)$, where $\Xi(z)$ is the adjoint polynomial matrix of $\Pi(z)$. Then, $\det(\Pi(B))\, u_t = \Xi(B)\, \epsilon_t$ and, by the definition of u_t in (3),

$$\Phi\left(B^{-1}\right) w_t = \Xi(B)\, \epsilon_t,$$

where, setting $\det(\Pi(z)) = a(z) = 1 - a_1 z - \cdots - a_{nr} z^{nr}$,

$$w_t = \det(\Pi(B))\, y_t = a(B) y_t. \tag{5}$$

Note that $\Xi(z)$ is a matrix polynomial of degree at most $(n-1)\,r$ and, because $\Pi(0) = I_n$, we also have $\Xi(0) = I_n$. By the latter condition in (2) one can find a

$0 < \delta_2 < 1$ such that $\Phi\left(z^{-1}\right)^{-1}\Xi\left(z\right)$ has a well defined power series representation

$$\Phi\left(z^{-1}\right)^{-1}\Xi\left(z\right) = \sum_{j=-(n-1)r}^{\infty} N_j z^{-j} = N\left(z^{-1}\right) \quad \text{for } |z| > 1-\delta_2. \tag{6}$$

Thus, the process w_t has the moving average representation

$$w_t = \sum_{j=-(n-1)r}^{\infty} N_j \epsilon_{t+j}, \tag{7}$$

where the coefficient matrices N_j decay to zero at a geometric rate as $j \to \infty$. Using the equalities in (6) one can solve these matrices recursively as functions of the parameters Π_j $(j = 1, ..., r)$ and Φ_j $(j = 1, ..., s)$ (see Appendix A.1). Finally, from (2) one obtains the moving average representation

$$y_t = \sum_{j=-\infty}^{\infty} \Psi_j \epsilon_{t-j}, \tag{8}$$

where Ψ_j $(n \times n)$ is the coefficient matrix of z^j in the Laurent series expansion of $\Psi\left(z\right) \overset{def}{=} \Phi\left(z^{-1}\right)^{-1}\Pi\left(z\right)^{-1}$ which exists for $1-\delta_2 < |z| < 1+\delta_1$ with Ψ_j decaying to zero at a geometric rate as $|j| \to \infty$. The representation (8) implies that y_t is a stationary and ergodic process with finite second moments.

Model (1) is referred to as the VAR(r, s) model. In the conventional causal case the abbreviation VAR(r) is also used. In the next section, we present the joint distribution of an observed time series generated by the VAR(r, s) process. This joint distribution is needed in the development of our forecasting methods and it also facilitates our discussion on parameter estimation and statistical inference.

2.2. Joint distribution of the VAR(r, s) process

As discussed in the Introduction, causal and noncausal autoregressions cannot be distinguished by second-order properties or the Gaussian likelihood. Therefore, it is necessary to assume that the error term ϵ_t is non-Gaussian. The theoretical results of Lanne and Saikkonen (2013) assume that the distribution of ϵ_t is of a fairly general elliptical form. However, an inspection of the arguments used in Section 3.1 of that paper reveals that this assumption is not needed to derive the distribution of the observed data and, therefore, it is not necessary for our

forecasting methods. Thus, unless otherwise indicated we only assume that the (non-Gaussian) distribution of ϵ_t is continuous with density function $f(\cdot)$, whose possible dependence on (unknown) parameters is not made explicit.

A detailed derivation of the joint distribution of the observed data can be found in Lanne and Saikkonen (2013), so here we only describe the final result. To this end, define the $n \times 1$ vectors

$$v_{k,T-s+k} = w_{T-s+k} - \sum_{j=-(n-1)r}^{-k} N_j \epsilon_{T-s+k+j}, \quad k = 1, ..., s, \tag{9}$$

where the sum is interpreted as zero when $k > (n-1)r$, that is, when the lower bound exceeds the upper bound (this convention will also be used elsewhere). Note also that, by (1) and (5), $v_{k,T-s+k}$ can be expressed as a function of the observed data $y_1, ..., y_T$ and that, by (7), the representation $v_{k,T-s+k} = \sum_{j=-k+1}^{\infty} N_j \epsilon_{T-s+k+j}$ holds, showing that $v_{k,T-s+k}$, $k = 1, ..., s$, are independent of ϵ_t, $t \leq T-s$. We also introduce the vector $\boldsymbol{z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3)$ where $\boldsymbol{z}_1 = (u_1, ..., u_r)$, $\boldsymbol{z}_2 = (\epsilon_{r+1}, ..., \epsilon_{T-s})$, and $\boldsymbol{z}_3 = (v_{1,T-s+1}, ..., v_{s,T})$ are independent in view of the preceding discussion and (4). These vectors can be expressed as functions of the observed data (and parameters), and in what follows we use a tilde to make this functional dependence explicit. Thus, the components of the vectors $\tilde{\boldsymbol{z}}_1$ and $\tilde{\boldsymbol{z}}_2$ are $\tilde{u}_t = \Phi(B^{-1}) y_t$, $t = 1, ..., r$, (see (3)) and $\tilde{\epsilon}_t = \Pi(B) \Phi(B^{-1}) y_t$, $t = r+1, ..., T-s$, (see (1)), respectively, whereas the components of $\tilde{\boldsymbol{z}}_3$, $\tilde{v}_{k,T-s+k}$, are defined by replacing w_{T-s+k} and $\epsilon_{T-s+k+j}$ on the right hand side of (9) by $a(B) y_{T-s+k}$ (see (5)) and $\tilde{\epsilon}_{T-s+k+j}$, $j = -(n-1)r, ..., -k$, $k = 1, ..., s$, respectively.

It is shown in Section 3.1 of Lanne and Saikkonen (2013) that the random vector $\boldsymbol{z}$ is related to the data vector $\boldsymbol{y} = (y_1, ..., y_T)$ according to $\boldsymbol{z} = \boldsymbol{H}_3 \boldsymbol{H}_2 \boldsymbol{H}_1 \boldsymbol{y}$, where $\boldsymbol{H}_1$, $\boldsymbol{H}_2$, and $\boldsymbol{H}_3$ $(T \times T)$ are nonsingular transformation matrices that depend on the parameters Π_j $(j = 1, ..., r)$ and Φ_j $(j = 1, ..., s)$ with $\boldsymbol{H}_2$ and $\boldsymbol{H}_3$ having unit determinant. Thus, it follows that the joint density function of the data vector $\boldsymbol{y}$ is given by (assuming T large enough)

$$p(\boldsymbol{y}) = h_{\boldsymbol{z}_1}(\tilde{\boldsymbol{z}}_1) \cdot \prod_{t=r+1}^{T-s} f(\tilde{\epsilon}_t) \cdot h_{\boldsymbol{z}_3}(\tilde{\boldsymbol{z}}_3) \cdot |\det(\boldsymbol{H}_1)|, \tag{10}$$

where $h_{\boldsymbol{z}_1}(\cdot)$ and $h_{\boldsymbol{z}_3}(\cdot)$ signify the density functions of the random vectors $\boldsymbol{z}_1$ and $\boldsymbol{z}_3$, respectively. For our subsequent developments the explicit expression of the matrix $\boldsymbol{H}_1$ is not relevant because the determinant term $|\det(\boldsymbol{H}_1)|$ will vanish from our forecasting formulas. In the purely noncausal case the joint density function $p(\boldsymbol{y})$ can be simplified by replacing the first factor $h_{\boldsymbol{z}_1}(\tilde{\boldsymbol{z}}_1)$ by unity, setting $r = 0$ and $\tilde{\epsilon}_t = \Phi(B^{-1})\, y_t$ in the second factor, and $\tilde{\boldsymbol{z}}_3 = (y_{T-s+1}, ..., y_T)$ in the third factor.

We shall now briefly discuss parameter estimation and statistical inference in the VAR(r, s) model (1). Following Lanne and Saikkonen (2013) we here assume that the error term ϵ_t has an elliptical distribution and use the second factor of the right hand side of (10) to obtain a computationally feasible approximation for the likelihood function. Maximizing this function over the permissible parameter space yields an (approximate) maximum likelihood (ML) estimator. Lanne and Saikkonen (2013) show that, under appropriate regularity conditions, the resulting (local) ML estimator is consistent and asymptotically normally distributed and that conventional methods to compute standard errors for estimators and to construct likelihood-based tests apply.

The preceding discussion assumes that the orders r and s of the VAR(r, s) model (1) are known. As in Lanne and Saikkonen (2013) we specify these orders as follows. First, using least squares or Gaussian ML we find a causal VAR(p) model that adequately describes the autocorrelation structure of the data with the order p determined by using conventional procedures such as model selection criteria and diagnostic checks. Then we check the residuals of this causal VAR(p) model for Gaussianity and, only when we detect deviations from Gaussianity, we consider noncausal VAR models. Next we choose a non-Gaussian error distribution, such as the multivariate t–distribution used in Lanne and Saikkonen (2013), and estimate all causal and noncausal VAR(r, s) models with the orders r and s summing to the selected order p. Finally, of these alternative models we choose the one that maximizes the likelihood function and evaluate its adequacy with conventional diagnostic tools.

3. Forecasting

In this section, we consider forecasting future observations y_{T+h} $(h \geq 1)$ and, unless otherwise stated, we shall assume that the model is not causal and not univariate, so that $s > 0$ and $n > 1$. We let $\mathsf{E}_T(\cdot)$ signify the conditional expectation operator given the observed data $\boldsymbol{y} = (y_1, ..., y_T)$.

Our starting point is equation (7) which we make operational by approximating the infinite sum therein by a finite sum. Specifically, from equations (5) and (7) we obtain the approximation

$$\mathsf{E}_T(y_{T+h}) \approx a_1 \mathsf{E}_T(y_{T+h-1}) + \cdots + a_{nr} \mathsf{E}_T(y_{T+h-nr}) + \mathsf{E}_T\left(\sum_{j=-(n-1)r}^{M-h} N_j \epsilon_{T+h+j}\right), \tag{11}$$

where $M > 0$ is supposed to be "large". As $\mathsf{E}_T(y_{T+h-j}) = y_{T+h-j}$ for $j \geq h$, (approximate) forecasts can be computed recursively starting from $h = 1$ if the last conditional expectation on the right hand side of (11) can be computed for every $h \geq 1$. In the univariate case $(n = 1)$ considered by Lanne et al. (2012b) this conditional expectation depends on the error terms $\epsilon_{T+1}, ..., \epsilon_{T+M}$ only. However, except for the purely noncausal case $(r = 0)$ this does not happen in our multivariate case, where the error terms $\epsilon_{T+1-(n-1)r}, ..., \epsilon_T$ are also involved and the fact that $\epsilon_{T-s+1}, ..., \epsilon_T$ $(s > 0)$ cannot be expressed as functions of the observed data (see (1)) causes complications. In the purely noncausal case these error terms vanish from the right hand side of (11), simplifying the situation and allowing a straightforward extension of the forecasting method of Lanne et al. (2012b). Therefore, and also to help understand the difficulties in the general case $(r > 0, s > 0)$, we shall first consider forecasting in the purely noncausal case. The general case requires a more delicate treatment provided in Section 3.2.

3.1. Purely noncausal case

In the purely noncausal case $(r = 0)$ the approximation (11) reduces to

$$\mathsf{E}_T(y_{T+h}) \approx \mathsf{E}_T\left(\sum_{j=0}^{M-h} N_j \epsilon_{T+h+j}\right), \quad N_0 = I_n. \tag{12}$$

To compute the conditional expectation on the right hand side we follow Lanne et al. (2012b) and derive the conditional density of $\boldsymbol{\epsilon}_+ = (\epsilon_{T+1}, ..., \epsilon_{T+M})$ given the data vector $\boldsymbol{y}$. Recall that now $\tilde{\epsilon}_t = \Phi\left(B^{-1}\right) y_t$ and $\tilde{\boldsymbol{z}}_3 = (y_{T-s+1}, ..., y_T)$. Using the expression of the density function $p\left(\boldsymbol{y}\right)$ in (10) and the preceding discussion one can check that the joint density function of $(\boldsymbol{y}, \boldsymbol{\epsilon}_+)$ can be written as

$$p\left(\boldsymbol{y}, \boldsymbol{\epsilon}_+\right) = \prod_{t=1}^{T-s} f\left(\tilde{\epsilon}_t\right) \cdot h_{\boldsymbol{z}_3, \boldsymbol{\epsilon}_+}(\boldsymbol{y}_3, \boldsymbol{\epsilon}_+) \cdot \left|\det\left(\boldsymbol{H}_1\right)\right|, \tag{13}$$

where $h_{\boldsymbol{z}_3, \boldsymbol{\epsilon}_+}(\boldsymbol{y}_3, \boldsymbol{\epsilon}_+)$ is the joint density function of $(\boldsymbol{z}_3, \boldsymbol{\epsilon}_+)$ and $\boldsymbol{y}_3 = (y_{T-s+1}, ..., y_T)$ (in this section we replace $\tilde{\boldsymbol{z}}_3$ by the more typical notation $\boldsymbol{y}_3$). From (10) (specialized to the present case) and (13) we find that the conditional density function of $\boldsymbol{\epsilon}_+$ given $\boldsymbol{y}$ is

$$p\left(\boldsymbol{\epsilon}_+ \mid \boldsymbol{y}\right) = \frac{h_{\boldsymbol{z}_3, \boldsymbol{\epsilon}_+}(\boldsymbol{y}_3, \boldsymbol{\epsilon}_+)}{h_{\boldsymbol{z}_3}(\boldsymbol{y}_3)} = \frac{h_{\boldsymbol{z}_3, \boldsymbol{\epsilon}_+}(\boldsymbol{y}_3, \boldsymbol{\epsilon}_+)}{\int h_{\boldsymbol{z}_3, \boldsymbol{\epsilon}_+}(\boldsymbol{y}_3, \boldsymbol{\epsilon}_+) d\boldsymbol{\epsilon}_+}.$$

The right hand side of (12) can thus be written as

$$\mathsf{E}_T\left(\sum_{j=0}^{M-h} N_j \epsilon_{T+h+j}\right) = \frac{\int \sum_{j=0}^{M-h} N_j \epsilon_{T+h+j} \cdot h_{\boldsymbol{z}_3, \boldsymbol{\epsilon}_+}(\boldsymbol{y}_3, \boldsymbol{\epsilon}_+) d\boldsymbol{\epsilon}_+}{\int h_{\boldsymbol{z}_3, \boldsymbol{\epsilon}_+}(\boldsymbol{y}_3, \boldsymbol{\epsilon}_+) d\boldsymbol{\epsilon}_+}. \tag{14}$$

As in Lanne et al. (2012b), we now derive a feasible approximation for the density function $h_{\boldsymbol{z}_3, \boldsymbol{\epsilon}_+}(\boldsymbol{y}_3, \boldsymbol{\epsilon}_+)$. As $y_t = \sum_{j=0}^{\infty} N_j \epsilon_{t+j}$ and $N_0 = I_n$, we have the approximate relation

$$\begin{bmatrix} I_n & N_1 & \cdots & \cdots & \cdots & \cdots & N_{M+s-1} \\ 0 & \ddots & \ddots & & & & \vdots \\ \vdots & \ddots & I_n & N_1 & \cdots & \cdots & N_M \\ \vdots & & \ddots & I_n & 0 & & 0 \\ \vdots & & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & I_n \end{bmatrix} \begin{bmatrix} \epsilon_{T-s+1} \\ \vdots \\ \epsilon_T \\ \epsilon_{T+1} \\ \vdots \\ \epsilon_{T+M} \end{bmatrix} \approx \begin{bmatrix} y_{T-s+1} \\ \vdots \\ y_T \\ \epsilon_{T+1} \\ \vdots \\ \epsilon_{T+M} \end{bmatrix},$$

or briefly $\boldsymbol{B}\boldsymbol{\epsilon}_{++} \approx \boldsymbol{v}$. As the matrix $\boldsymbol{B}$ is nonsingular with unit determinant this yields $\boldsymbol{\epsilon}_{++} \approx \boldsymbol{B}^{-1}\boldsymbol{v}$ or

$$\left(\epsilon_{T-s+1}, ..., \epsilon_T, \epsilon_{T+1}, ..., \epsilon_{T+M}\right) \approx \left(\tilde{\epsilon}_{T-s+1}\left(\boldsymbol{\epsilon}_+\right), ..., \tilde{\epsilon}_T\left(\boldsymbol{\epsilon}_+\right), \epsilon_{T+1}, ..., \epsilon_{T+M}\right),$$

where $\tilde{\epsilon}_{T-s+1}(\boldsymbol{\epsilon}_+), ..., \tilde{\epsilon}_T(\boldsymbol{\epsilon}_+)$ $(n \times 1)$ are the first s (vector) components of the vector $\boldsymbol{B}^{-1}\boldsymbol{v}$, and hence dependent on $y_{T-s+1}, ..., y_T$. Thus, it follows that the density function $h_{\boldsymbol{z}_3,\boldsymbol{\epsilon}_+}(\boldsymbol{y}_3, \boldsymbol{\epsilon}_+)$ can be approximated as

$$h_{\boldsymbol{z}_3,\boldsymbol{\epsilon}_+}(\boldsymbol{y}_3, \boldsymbol{\epsilon}_+) \approx \prod_{j=1}^{s} f\left(\tilde{\epsilon}_{T-s+j}(\boldsymbol{\epsilon}_+)\right) \cdot \prod_{t=T+1}^{T+M} f(\epsilon_t). \tag{15}$$

As in Lanne et al. (2012b), we can use this approximation to compute approximations for the two integrals on the right hand side of (14). More generally, for any function of $\boldsymbol{\epsilon}_+$, say $q(\boldsymbol{\epsilon}_+)$, we can use (15) to obtain

$$\mathsf{E}_T\left(q(\boldsymbol{\epsilon}_+)\right) \approx \frac{\int q(\boldsymbol{\epsilon}_+) \cdot \prod_{j=1}^{s} f\left(\tilde{\epsilon}_{T-s+j}(\boldsymbol{\epsilon}_+)\right) \cdot \prod_{t=T+1}^{T+M} f(\epsilon_t)\, d\boldsymbol{\epsilon}_+}{\int \prod_{j=1}^{s} f\left(\tilde{\epsilon}_{T-s+j}(\boldsymbol{\epsilon}_+)\right) \cdot \prod_{t=T+1}^{T+M} f(\epsilon_t)\, d\boldsymbol{\epsilon}_+}.$$

(Here as well as in similar subsequent instances existence and finiteness of the stated expectations are assumed.) The numerator on the right hand side can be interpreted as the expectation of the product of the first two factors in the integrand with respect to the distribution of $\boldsymbol{\epsilon}_+ = (\epsilon_{T+1}, ..., \epsilon_{T+M})$, whereas the denominator can be interpreted as the expectation of $\prod_{j=1}^{s} f\left(\tilde{\epsilon}_{T-s+j}(\boldsymbol{\epsilon}_+)\right)$ with respect to the same distribution. Using Monte Carlo simulation, we can therefore approximate $\mathsf{E}_T\left(q(\boldsymbol{\epsilon}_+)\right)$ by

$$\hat{\mathsf{E}}_T\left(q(\boldsymbol{\epsilon}_+)\right) = \frac{\frac{1}{m}\sum_{i=1}^{m} q(\boldsymbol{\epsilon}_+^{(i)}) \cdot \prod_{j=1}^{s} f(\tilde{\epsilon}_{T-s+j}(\boldsymbol{\epsilon}_+^{(i)})}{\frac{1}{m}\sum_{i=1}^{m} \prod_{j=1}^{s} f(\tilde{\epsilon}_{T-s+j}(\boldsymbol{\epsilon}_+^{(i)}))}, \tag{16}$$

where $\boldsymbol{\epsilon}_+^{(i)} = (\epsilon_{T+1}^{(i)}, ..., \epsilon_{T+M}^{(i)})$, $i = 1, ..., m$, are mutually independent simulated realizations from the distribution of $\boldsymbol{\epsilon}_+$ so that $\epsilon_{T+1}^{(i)}, ..., \epsilon_{T+M}^{(i)}$ are independent random vectors for every i. As $m \to \infty$, the right hand side of (16) converges almost surely and provides an approximation for $\mathsf{E}_T\left(q(\boldsymbol{\epsilon}_+)\right)$ that can be made arbitrarily accurate by choosing m and M large enough.

To obtain forecasts for y_{T+h} $(h \geq 1)$ one needs to compute values of the right hand side of (16) with $q(\boldsymbol{\epsilon}_+) = \sum_{j=0}^{M-h} N_j \epsilon_{T+h+j}$ (see (14)). Specifically, we have the following forecasting procedure.

Step 1. Generate $\boldsymbol{\epsilon}_+^{(i)} = (\epsilon_{T+1}^{(i)}, ..., \epsilon_{T+M}^{(i)})$, $i = 1, ..., m$, as described below (16).

Step 2. Compute the forecasts $\hat{\mathsf{E}}_T(y_{T+h})$, $h = 1, 2, ...$, by choosing $q(\boldsymbol{\epsilon}_+) = \sum_{j=0}^{M-h} N_j \epsilon_{T+h+j}$ in (16).

For m and M large enough, the resulting forecasts approximate the true forecast $\mathsf{E}_T(y_{T+h})$ arbitrarily closely. Appendix A.1 shows how to compute the coefficient matrices N_j recursively as functions of the parameters Π_j $(j = 1, ..., r)$ and Φ_j $(j = 1, ..., s)$. Choosing the values of the integers m and M will be discussed in Section 4.

Proceeding as in Lanne et al. (2012b) we can also obtain interval forecasts and forecasts for the conditional distribution of the components of y_{T+h} $(h \geq 1)$. Let $\mathbf{1}(\cdot)$ stand for the indicator function and $\iota_a = (0, ..., 0, 1, 0, ..., 0)$ $(n \times 1)$ the ath unit vector. Then a forecast for the conditional cumulative distribution function of $y_{a,T+h} = \iota_a' y_{T+h}$, the ath component of y_{T+h}, at point $x \in \mathbb{R}$ is obtained as (see (5) and (7))

$$\mathsf{E}_T\left(\mathbf{1}\left(y_{a,T+h} \leq x\right)\right) \approx \mathsf{E}_T\left(\mathbf{1}\left(\sum_{j=0}^{M-h} \iota_a' N_j \epsilon_{T+h+j} \leq x\right)\right),$$

where the right hand side can be approximated by using (16) with $q(\boldsymbol{\epsilon}_+) = \mathbf{1}\left(\sum_{j=0}^{M-h} \iota_a' N_j \epsilon_{T+h+j} \leq x\right)$. Thus, choosing a grid $x_1, ..., x_K$ with a large enough value of K, one can obtain a forecast of the whole conditional cumulative distribution function of $y_{a,T+h}$, and using appropriate quantiles from the lower and upper tails of this forecast an interval forecast for $y_{a,T+h}$ can be constructed for any $h \geq 1$.

3.2. General case

As already indicated, the general noncausal case seems to require techniques more burdensome than those in the purely noncausal case (or in the general univariate noncausal case). To demonstrate this, consider the joint density of the augmented data vector $(\boldsymbol{y}, \boldsymbol{\epsilon}_+)$ and conclude from the discussion leading to the density function $p(\boldsymbol{y})$ in (10) that the joint density of $(\boldsymbol{y}, \boldsymbol{\epsilon}_+)$, and hence the conditional density of $\boldsymbol{\epsilon}_+$ given $\boldsymbol{y}$, involves the joint density of $(\boldsymbol{z}_3, \boldsymbol{\epsilon}_+)$. For simplicity, suppose that $s = 1$ so that $\boldsymbol{z}_3 = v_{1,T} = \sum_{j=0}^{\infty} N_j \epsilon_{T+j}$ and $\boldsymbol{z}_3 \approx \sum_{j=0}^{M} N_j \epsilon_{T+j}$ for M large (see (9) and the subsequent discussion). In the purely noncausal case we have $N_0 = I_n$, but this does not hold in the general case and it is even possible

that the matrix N_0 is singular. This happens, for example, when $r = s = 1$ and

$$\Pi_1 = \begin{bmatrix} 0 & 0 \\ -3/4 & 3/4 \end{bmatrix} \text{ and } \Phi_1 = \begin{bmatrix} 2/3 & 2/3 \\ 0 & 0 \end{bmatrix}.$$

When the matrix N_0 is singular the random vectors $\boldsymbol{z}_3$ and $\boldsymbol{\epsilon}_+ = (\epsilon_{T+1}, ..., \epsilon_{T+M})$ are approximately linearly dependent so that, apart from the approximation error, the joint distribution of $\boldsymbol{z}_3$ and $\boldsymbol{\epsilon}_+$ is singular. This makes the conventional use of the joint density of $\boldsymbol{z}_3$ and $\boldsymbol{\epsilon}_+$, employed in the purely noncausal case, inappropriate.

To overcome the difficulty described above we first develop a procedure that is generally applicable but requires the use of importance sampling not needed in the purely noncausal case considered in the preceding section. In Section 3.2.2, we show how a simpler technique, similar to that derived in the purely noncausal case, can be obtained when a suitable condition about the structure of the model holds. When $s = 1$ this condition requires that the matrix N_0 is nonsingular.

3.2.1. Importance-sampling-based forecasting

For our subsequent discussion it appears convenient to write the approximate forecasting formula (11) as

$$\begin{aligned} \mathsf{E}_T\left(y_{T+h}\right) \approx\ & a_1\mathsf{E}_T\left(y_{T+h-1}\right) + \cdots + a_{nr}\mathsf{E}_T\left(y_{T+h-nr}\right) + \sum_{j=-(n-1)r}^{-s-h} N_j\tilde{\epsilon}_{T+h+j} \\ & + \mathsf{E}_T\left(\sum_{j=-s-h+1}^{-s-h+sn} N_j\epsilon_{T+h+j}\right) + \mathsf{E}_T\left(\sum_{j=-s-h+sn+1}^{M-h} N_j\epsilon_{T+h+j}\right), \quad (17) \end{aligned}$$

where we have divided the sum involving the error terms into three components of which the first one depends on the data and the second one contains the error terms that will be treated in a special manner.

Our subsequent developments make use of the $sn \times n$ matrices

$$\boldsymbol{N}_j = \begin{bmatrix} N_j \\ \vdots \\ N_{j-s+1} \end{bmatrix}, \quad j = 0, 1, ...$$

It is demonstrated in Appendix A.1 that the matrix $[\boldsymbol{N}_0 \ \cdots \ \boldsymbol{N}_{sn-1}]$ $(sn \times sn^2)$ is of full row rank, implying that we can find a matrix $[\boldsymbol{K}_0 \ \cdots \ \boldsymbol{K}_{sn-1}]$ $(sn\,(n-1) \times sn^2)$ such that the matrix

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{N}_0 \ \cdots \ \boldsymbol{N}_{sn-1} \\ \boldsymbol{K}_0 \ \cdots \ \boldsymbol{K}_{sn-1} \end{bmatrix}, \quad sn^2 \times sn^2, \tag{18}$$

is nonsingular. One possibility that always applies is to choose the rows of $[\boldsymbol{K}_0 \ \cdots \ \boldsymbol{K}_{sn-1}]$ as basis vectors of the orthogonal complement of the space spanned by the rows of $[\boldsymbol{N}_0 \ \cdots \ \boldsymbol{N}_{sn-1}]$. A simpler choice that applies when the matrix $[\boldsymbol{N}_0 \ \cdots \ \boldsymbol{N}_{s-1}]$ $(sn \times sn)$ is nonsingular will be discussed in the next section. Using the matrix $\boldsymbol{Q}$ introduced in (18) we define the vector

$$\begin{bmatrix} \boldsymbol{\zeta}_1 \\ \boldsymbol{\zeta}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{N}_0 \ \cdots \ \boldsymbol{N}_{sn-1} \\ \boldsymbol{K}_0 \ \cdots \ \boldsymbol{K}_{sn-1} \end{bmatrix} \begin{bmatrix} \epsilon_{T-s+1} \\ \vdots \\ \epsilon_{T-s+sn} \end{bmatrix}, \quad sn^2 \times 1, \tag{19}$$

where $\boldsymbol{\zeta}_1$ is $sn \times 1$, $\boldsymbol{\zeta}_2$ is $sn\,(n-1) \times 1$, and the error terms on the right hand side are the ones in the penultimate term on the right hand side of (17). Furthermore, as $\boldsymbol{z}_3 = (v_{1,T-s+1}, ..., v_{s,T})$ with $v_{k,T-s+k} = \sum_{j=-k+1}^{\infty} N_j \epsilon_{T-s+k+j}$ (see the discussion following (9)) the definition of $\boldsymbol{N}_j$ shows that $\boldsymbol{z}_3 = \sum_{j=0}^{\infty} \boldsymbol{N}_j \epsilon_{T-s+1+j}$. Hence, we have $\boldsymbol{\zeta}_1 = \boldsymbol{z}_3 - \sum_{j=sn}^{\infty} \boldsymbol{N}_j \epsilon_{T-s+j+1}$, which will be used below.

Now, use equations (18) and (19) to write the sum in the penultimate term on the right hand side of (17) as

$$\sum_{j=-s-h+1}^{-s-h+sn} N_j \epsilon_{T+h+j} = \boldsymbol{P}_h \begin{bmatrix} \boldsymbol{\zeta}_1 \\ \boldsymbol{\zeta}_2 \end{bmatrix},$$

where $\boldsymbol{P}_h = [N_{-s-h+1} \ \cdots \ N_{-s-h+sn}]\,\boldsymbol{Q}^{-1}$. Thus, using the approximation $\boldsymbol{\zeta}_1 \approx \boldsymbol{z}_3 - \sum_{j=sn}^{M+s-1} \boldsymbol{N}_j \epsilon_{T-s+j+1}$ we obtain

$$\mathsf{E}_T \left(\sum_{j=-s-h+1}^{-s-h+sn} N_j \epsilon_{T+h+j} \right) \approx \boldsymbol{P}_h \begin{bmatrix} \mathsf{E}_T \left(\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+) \right) \\ \mathsf{E}_T(\boldsymbol{\zeta}_2) \end{bmatrix}, \tag{20}$$

where $\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+) = \tilde{\boldsymbol{z}}_3 - \sum_{j=sn}^{M+s-1} \boldsymbol{N}_j \epsilon_{T-s+j+1}$ with $\boldsymbol{e}_+ = (\epsilon_{T-s+sn+1}, ..., \epsilon_{T+M})$ and $\boldsymbol{\zeta}_2 = \sum_{j=0}^{sn-1} \boldsymbol{K}_j \epsilon_{T-s+j+1}$ (see (19)). From this and the approximation (17) it follows

that to obtain forecasts for y_{T+h} $(h \geq 1)$ we should be able to obtain forecasts for (functions of) $\boldsymbol{e}_+$ and $\boldsymbol{\zeta}_2$. To this end, we consider the extended data vector $(\boldsymbol{y}, \boldsymbol{\zeta}_2, \boldsymbol{e}_+)$ and derive the conditional density of $(\boldsymbol{\zeta}_2, \boldsymbol{e}_+)$ given $\boldsymbol{y}$.

It is shown in Appendix A.2 that, for M large, the conditional density of $(\boldsymbol{\zeta}_2, \boldsymbol{e}_+)$ given $\boldsymbol{y}$ can be approximated by using the joint density of the independent random vectors $(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2)$ and $\boldsymbol{e}_+$ or, specifically,

$$p\left((\boldsymbol{\zeta}_2, \boldsymbol{e}_+) \mid \boldsymbol{y}\right) \approx \frac{h_{\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2}(\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+), \boldsymbol{\zeta}_2) \cdot \prod_{t=T-s+sn+1}^{T+M} f\left(\epsilon_t\right)}{\int\int h_{\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2}(\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+), \boldsymbol{\zeta}_2) \cdot \prod_{t=T-s+sn+1}^{T+M} f\left(\epsilon_t\right) d\boldsymbol{\zeta}_2 d\boldsymbol{e}_+}. \tag{21}$$

Here the notation is as follows. First, $\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+)$ is as in (20) with the $(n \times 1$ vector) components $\tilde{\zeta}_{1,k}(\boldsymbol{e}_+) = \tilde{v}_{k,T-s+k} - \sum_{j=sn}^{M+s-1} N_{j-k+1}\epsilon_{T-s+j+1}$ $(k = 1, ..., s)$. Second, $h_{\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2}(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2)$ is the joint density function of $\boldsymbol{\zeta}_1$ and $\boldsymbol{\zeta}_2$, and defined as

$$h_{\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2}(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2) = \prod_{j=1}^{sn} f\left(\sum_{k=1}^{s} R_{j,k}\zeta_{1,k} + \sum_{k=s+1}^{sn} R_{j,k}\zeta_{2,k}\right) \cdot \left|\det\left(\boldsymbol{R}\right)\right|, \tag{22}$$

where $\zeta_{1,k}$ and $\zeta_{2,k}$ signify the kth $(n \times 1$ vector) components of $\boldsymbol{\zeta}_1$ and $\boldsymbol{\zeta}_2$, and $\boldsymbol{R} = [R_{j,k}] = \boldsymbol{Q}^{-1}$ $(j, k = 1, ..., n)$ with the partitions $R_{j,k}$ of order $n \times n$.

Now, as discussed below (20), to obtain forecasts for y_{T+h} $(h \geq 1)$ we should be able to compute (an approximation for) the conditional expectation $\mathsf{E}_T(q(\boldsymbol{\zeta}_2, \boldsymbol{e}_+))$ with $q(\boldsymbol{\zeta}_2, \boldsymbol{e}_+)$ a function of $q(\boldsymbol{\zeta}_2, \boldsymbol{e}_+)$. From (21) we find that

$$\mathsf{E}_T(q(\boldsymbol{\zeta}_2, \boldsymbol{e}_+)) \approx \frac{\int\int q(\boldsymbol{\zeta}_2, \boldsymbol{e}_+) \cdot h_{\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2}(\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+), \boldsymbol{\zeta}_2) \cdot \prod_{t=T-s+sn+1}^{T+M} f\left(\epsilon_t\right) d\boldsymbol{\zeta}_2 d\boldsymbol{e}_+}{\int\int h_{\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2}(\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+), \boldsymbol{\zeta}_2) \cdot \prod_{t=T-s+sn+1}^{T+M} f\left(\epsilon_t\right) d\boldsymbol{\zeta}_2 d\boldsymbol{e}_+}, \tag{23}$$

where $h_{\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2}(\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+)$ is obtained from (22) by replacing $\zeta_{1,k}$ by $\tilde{\zeta}_{1,k}(\boldsymbol{e}_+)$.

Numerical solutions for the integrals on the right hand side of (23) can be obtained but techniques more complicated than in the preceding section or in Lanne et al. (2012b) seem to be required. As in Breidt and Hsu (2005), where an analogous forecasting procedure for (univariate) noninvertible moving average models is developed, one can employ an importance sampling technique (see, e.g., Sec. 4.3 of Geweke (1996)). To this end, let $\varphi\left(\cdot\right)$ be an $sn\left(n-1\right)$-dimensional density function whose support contains the support of the distribution of $\boldsymbol{\zeta}_2$, and define

$$W(\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+), \boldsymbol{\zeta}_2) = \frac{h_{\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2}(\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+), \boldsymbol{\zeta}_2)}{\varphi(\boldsymbol{\zeta}_2)}. \tag{24}$$

Then, the numerator in (23) can be written as

$$\int\int q(\boldsymbol{\zeta}_2, \boldsymbol{e}_+) \cdot W(\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+), \boldsymbol{\zeta}_2) \cdot \varphi(\boldsymbol{\zeta}_2) \cdot \prod_{t=T-s+sn+1}^{T+M} f(\epsilon_t)\, d\boldsymbol{\zeta}_2 d\boldsymbol{e}_+ \tag{25}$$

and the denominator in (23) can similarly be written as

$$\int\int W(\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+), \boldsymbol{\zeta}_2) \cdot \varphi(\boldsymbol{\zeta}_2) \cdot \prod_{t=T-s+sn+1}^{T+M} f(\epsilon_t)\, d\boldsymbol{\zeta}_2 d\boldsymbol{e}_+. \tag{26}$$

Clearly, the integral in (25) is the expectation of $q(\boldsymbol{\zeta}_2, \boldsymbol{e}_+) \cdot W(\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+), \boldsymbol{\zeta}_2)$ with respect to a distribution with density $\varphi \times f \times \cdots \times f$ ($M - sn + s$ copies of f) and the integral in (26) is the expectation of $W(\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+), \boldsymbol{\zeta}_2)$ with respect to the same distribution. Thus, the conditional expectation in (23) can be approximated via Monte Carlo simulation as

$$\hat{\mathsf{E}}_T(q(\boldsymbol{\zeta}_2, \boldsymbol{e}_+)) = \frac{\frac{1}{m}\sum_{i=1}^{m} q(\boldsymbol{\zeta}_2^{(i)}, \boldsymbol{e}_+^{(i)}) \cdot W(\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+^{(i)}), \boldsymbol{\zeta}_2^{(i)})}{\frac{1}{m}\sum_{i=1}^{m} W(\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+^{(i)}), \boldsymbol{\zeta}_2^{(i)})}, \tag{27}$$

where $(\boldsymbol{\zeta}_2^{(i)}, \boldsymbol{e}_+^{(i)}) = (\boldsymbol{\zeta}_2^{(i)}, \epsilon_{T-s+sn+1}^{(i)}, ..., \epsilon_{T+M}^{(i)})$, $i = 1, ..., m$, are mutually independent simulated realizations from a distribution with density $\varphi \times f \times \cdots \times f$ (regarding $\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+^{(i)})$, see equation (20)). Thus, $\boldsymbol{\zeta}_2^{(i)}$ ($sn\,(n-1) \times 1$) is drawn from a distribution with density φ and $\epsilon_{T-s+sn+1}^{(i)}, ..., \epsilon_{T+M}^{(i)}$ ($n \times 1$) are drawn independently of $\boldsymbol{\zeta}_2^{(i)}$ from a distribution with density f and, similarly to $\epsilon_{T-s+sn+1}, ..., \epsilon_{T+M}$, the random vectors $\epsilon_{T-s+sn+1}^{(i)}, ..., \epsilon_{T+M}^{(i)}$ are independent for every i. Finally, $W(\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+^{(i)}), \boldsymbol{\zeta}_2^{(i)})$ is computed by using (22) and (24).

Approximate forecasts, which can be made arbitrarily accurate by choosing m and M large enough, can be obtained recursively as follows.

Step 1. Generate $(\boldsymbol{\zeta}_2^{(i)}, \boldsymbol{e}_+^{(i)}) = (\boldsymbol{\zeta}_2^{(i)}, \epsilon_{T-s+sn+1}^{(i)}, ..., \epsilon_{T+M}^{(i)})$, $i = 1, ..., m$, as described below (27).

Step 2. For $h = 1, 2, ...$, apply (27) recursively with (see (17) and (20))

$$q(\boldsymbol{\zeta}_2, \boldsymbol{e}_+) = \boldsymbol{P}_h \begin{bmatrix} \tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+) \\ \boldsymbol{\zeta}_2 \end{bmatrix} + \sum_{j=-s-h+sn+1}^{M-h} N_j \epsilon_{T+h+j} \stackrel{def}{=} q_h(\boldsymbol{\zeta}_2, \boldsymbol{e}_+),$$

and compute

$$\hat{\mathsf{E}}_T(y_{T+h}) = a_1 \hat{\mathsf{E}}_T(y_{T+h-1}) + \cdots + a_{nr} \hat{\mathsf{E}}_T(y_{T+h-nr}) + \sum_{j=-(n-1)r}^{-s-h} N_j \tilde{\epsilon}_{T+h+j} + \hat{\mathsf{E}}_T(q_h(\boldsymbol{\zeta}_2, \boldsymbol{e}_+)),$$

where $\hat{\mathsf{E}}_T(y_{T+h-k}) = y_{T+h-k}$ for $k \geq h$ and $N_j = 0$ for $j < -(n-1)r$. Thus, $\sum_{j=-(n-1)r}^{-s-h} N_j \tilde{\epsilon}_{T+h+j} = 0$ for $s + h > (n-1)r$ and the first term in the definition of $q_h(\boldsymbol{\zeta}_2, \boldsymbol{e}_+)$ vanishes when $h + s - sn > (n-1)r$.

In addition to choosing values for the integers m and M (to be discussed in Section 4) the application of the preceding procedure requires two choices. First, one has to choose the matrix $[\boldsymbol{K}_0 \ \cdots \ \boldsymbol{K}_{sn-1}]$ $(sn(n-1) \times sn^2)$ whose rows we here assume to be formed of the basis vectors of the orthogonal complement of the space spanned by the rows of $[\boldsymbol{N}_0 \ \cdots \ \boldsymbol{N}_{sn-1}]$. Second, one has to choose the $sn(n-1)$-dimensional auxiliary density function $\varphi(\boldsymbol{\zeta}_2)$. As $\boldsymbol{\zeta}_2 = \sum_{j=0}^{sn-1} \boldsymbol{K}_j \epsilon_{T-s+1+j}$, a potentially reasonable choice might be based on the chosen error distribution. In the bivariate special case with $s = 1$ the random vector $\boldsymbol{\zeta}_2$ is also bivariate, and one could choose $\varphi(\boldsymbol{\zeta}_2)$ as the density function of the error term ϵ_t. In general, as the dimension of $\boldsymbol{\zeta}_2$ is $s(n-1)$ times the dimension of ϵ_t, one could similarly choose $\varphi(\boldsymbol{\zeta}_2)$ as the density function of $(\epsilon_{T-s+1}, ..., \epsilon_{T-s+sn})$, that is, $f \times \cdots \times f$ $(sn(n-1)$ copies). This choice is probably not optimal but, due to its simplicity, will be used in our subsequent numerical illustrations where the error term is assumed to have a multivariate t–distribution. Breidt and Hsu (2005) use a somewhat similar importance sampler in their forecasting procedure.

As in the purely noncausal case, it is also possible to obtain interval forecasts and forecasts for the conditional distribution of the components of y_{T+h} $(h \geq 1)$. We illustrate this below in the case of one-step-ahead forecasts $(h = 1)$ and provide details of the more complex general case $(h \geq 1)$ in Appendix A.4. Using the notation introduced at the end of Section 3.1 the optimal forecast for the conditional cumulative distribution function of the ath component of y_{T+1}, at point $x \in \mathbb{R}$ is (see (5) and (7))

$$\mathsf{E}_T\left(\mathbf{1}\left(y_{a,T+1} \leq x\right)\right) \approx \mathsf{E}_T\left(\mathbf{1}\left(\sum_{j=1}^{nr} a_j \iota_a' y_{T+1-j} + \sum_{j=-(n-1)r}^{M-1} \iota_a' N_j \epsilon_{T+1+j} \leq x\right)\right).$$

Decomposing the latter sum inside the indicator function as in (17) we have

$$\begin{aligned}\mathsf{E}_T\left(\mathbf{1}\left(y_{a,T+1}\le x\right)\right) &\approx \mathsf{E}_T\left(\mathbf{1}\left(\sum_{j=-s}^{-s-1+sn}\iota_a' N_j\epsilon_{T+1+j}+\sum_{j=-s+sn}^{M-1}\iota_a' N_j\epsilon_{T+1+j}\le x-\iota_a'\tilde{\kappa}_{T,1}\right)\right)\\ &\approx \mathsf{E}_T\left(\mathbf{1}\left(\iota_a'\boldsymbol{P}_1\begin{bmatrix}\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+)\\ \boldsymbol{\zeta}_2\end{bmatrix}+\sum_{j=-s+sn}^{M-1}\iota_a' N_j\epsilon_{T+1+j}\le x-\iota_a'\tilde{\kappa}_{T,1}\right)\right),\end{aligned}$$

where the latter approximation is based on the discussion leading to (20) and, for brevity,

$$\tilde{\kappa}_{T,1}=\sum_{j=1}^{nr}a_j y_{T+1-j}+\sum_{j=-(n-1)r}^{-s-1}N_j\tilde{\epsilon}_{T+1+j}.$$

Note that $\tilde{\kappa}_{T,1}$ depends on the observed data and is treated as fixed, and the same applies to $\tilde{\boldsymbol{z}}_3$ which appears in the vector $\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+)$ (see (20)). Thus, to obtain (an approximation for) $\mathsf{E}_T\left(\mathbf{1}\left(y_{a,T+1}\le x\right)\right)$ we need to compute the conditional expectation of $\mathsf{E}_T\left(q(\boldsymbol{\zeta}_2,\boldsymbol{e}_+)\right)$ with

$$q(\boldsymbol{\zeta}_2,\boldsymbol{e}_+)=\mathbf{1}\left(\iota_a'\boldsymbol{P}_1\begin{bmatrix}\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+)\\ \boldsymbol{\zeta}_2\end{bmatrix}+\sum_{j=-s+sn}^{M-1}\iota_a' N_j\epsilon_{T+1+j}\le x-\iota_a'\tilde{\kappa}_{T,1}\right). \qquad (28)$$

Using this choice of $q(\boldsymbol{\zeta}_2,\boldsymbol{e}_+)$ in (27) and the subsequent Steps 1 and 2 yields a forecast for the conditional cumulative distribution function of $y_{a,T+1}$ at point x. A forecast of the whole conditional cumulative distribution function and interval forecast for $y_{a,T+1}$ can be obtained as described at the end of the preceding section.

3.2.2. Forecasting without importance sampling

It is possible to simplify the preceding simulation method if suitable knowledge of the structure of the matrix $[\boldsymbol{N}_0\ \cdots\ \boldsymbol{N}_{sn-1}]$ is available. In particular, as will be seen below, it is possible to avoid the use of importance sampling if the matrix $[\boldsymbol{N}_0\ \cdots\ \boldsymbol{N}_{s-1}]$ $(sn\times sn)$ is nonsingular, for then we can choose

$$\boldsymbol{Q}=\begin{bmatrix}\boldsymbol{N}_0&\cdots&\boldsymbol{N}_{sn-1}\\ \boldsymbol{K}_0&\cdots&\boldsymbol{K}_{sn-1}\end{bmatrix}\overset{def}{=}\begin{bmatrix}\boldsymbol{Q}_{11}&\boldsymbol{Q}_{12}\\ 0&I_{sn(n-1)}\end{bmatrix},$$

where $\boldsymbol{Q}_{11}=[\boldsymbol{N}_0\ \cdots\ \boldsymbol{N}_{s-1}]$, $\boldsymbol{Q}_{12}=[\boldsymbol{N}_s\ \cdots\ \boldsymbol{N}_{sn-1}]$ and $[\boldsymbol{K}_0\ \cdots\ \boldsymbol{K}_{sn-1}]=\left[0:I_{sn(n-1)}\right]$. In the purely noncausal case considered in Section 3.1, this choice is always possible because then $N_j=0$, $j<0$, and $N_0=I_n$, implying that the

matrix $\boldsymbol{Q}_{11}$ is upper triangular with unit diagonal elements. However, in general the preceding choice of $[\boldsymbol{K}_0 \ \cdots \ \boldsymbol{K}_{sn-1}]$ may be inappropriate because the nonsingularity of the matrix $\boldsymbol{Q}_{11}$ may fail (see the example at the beginning of Section 3.2 where $s = 1$ and $\boldsymbol{Q}_{11} = \boldsymbol{N}_0 = N_0$ is singular). On the other hand, in practice the matrix $\boldsymbol{Q}_{11}$ is unknown and has to replaced by an estimate which, due to estimation errors, is necessarily nonsingular (with probability one). Moreover, a simulated example provided in the next section suggests that, even when the assumed nonsingularity does not hold, the forecasting procedure to be derived in this section performs well compared to its robust but computationally more demanding alternative developed in the previous section. Note also that in practice one can assess the possible singularity of $\boldsymbol{Q}_{11}$ by examining, for example, the eigenvalues or determinant of its estimate.

When the matrix $\boldsymbol{Q}$ is as defined above, we have $\boldsymbol{\zeta}_2 = (\epsilon_{T+1}, ..., \epsilon_{T-s+sn})$ (see (19)) and $\boldsymbol{R} = \boldsymbol{Q}^{-1}$ is of the same form as $\boldsymbol{Q}$ or, specifically,

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ 0 & I_{sn(n-1)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Q}_{11}^{-1} & -\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12} \\ 0 & I_{sn(n-1)} \end{bmatrix}.$$

Thus, in this case the joint density function of $\boldsymbol{\zeta}_1$ and $\boldsymbol{\zeta}_2$ becomes (see (22))

$$h_{\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2}(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2) = \prod_{j=1}^{s} f\left(\sum_{k=1}^{s} R_{j,k}\zeta_{1,k} + \sum_{k=s+1}^{sn} R_{j,k}\epsilon_{T-s+k}\right) \cdot \prod_{j=s+1}^{sn} f\left(\epsilon_{T-s+j}\right) \cdot \left|\det\left(\boldsymbol{R}\right)\right|,$$

so that the approximate relation (23) can be written as

$$\mathsf{E}_T\left(q(\boldsymbol{\zeta}_2, \boldsymbol{e}_+)\right) \approx \frac{\int\int q\left(\boldsymbol{\zeta}_2, \boldsymbol{e}_+\right) \cdot \prod_{j=1}^{s} f\left(\sum_{k=1}^{s} R_{j,k}\tilde{\zeta}_{1,k}(\boldsymbol{e}_+) + \sum_{k=s+1}^{sn} R_{j,k}\epsilon_{T-s+k}\right) \cdot \prod_{t=T+1}^{T+M} f\left(\epsilon_t\right) d\boldsymbol{\zeta}_2 d\boldsymbol{e}_+}{\int\int \prod_{j=1}^{s} f\left(\sum_{k=1}^{s} R_{j,k}\tilde{\zeta}_{1,k}(\boldsymbol{e}_+) + \sum_{k=s+1}^{sn} R_{j,k}\epsilon_{T-s+k}\right) \cdot \prod_{t=T+1}^{T+M} f\left(\epsilon_t\right) d\boldsymbol{\zeta}_2 d\boldsymbol{e}_+},$$

where $\tilde{\zeta}_{1,k}(\boldsymbol{e}_+)$ is defined below (21). Thus, as now $(\boldsymbol{\zeta}_2, \boldsymbol{e}_+) = (\epsilon_{T+1}, ..., \epsilon_{T+M})$, the integral in the numerator is the expectation of

$$q\left(\boldsymbol{\zeta}_2, \boldsymbol{e}_+\right) \cdot \prod_{j=1}^{s} f\left(\sum_{k=1}^{s} R_{j,k}\tilde{\zeta}_{1,k}(\boldsymbol{e}_+) + \sum_{k=s+1}^{sn} R_{j,k}\epsilon_{T-s+k}\right)$$

with respect to a distribution with density $f \times \cdots \times f$ (M copies) whereas the

integral in the denominator is the expectation of

$$\prod_{j=1}^{s} f\left(\sum_{k=1}^{s} R_{j,k}\tilde{\zeta}_{1,k}(\boldsymbol{e}_+) + \sum_{k=s+1}^{sn} R_{j,k}\epsilon_{T-s+k}\right)$$

with respect to the same distribution.

The preceding discussion shows that, instead of (27), we can approximate the conditional expectation $\mathsf{E}_T\left(q\left(\boldsymbol{\zeta}_2, \boldsymbol{e}_+\right)\right)$ via Monte Carlo simulation as

$$\hat{\mathsf{E}}_T\left(q\left(\boldsymbol{\zeta}_2, \boldsymbol{e}_+\right)\right) = \frac{\frac{1}{m}\sum_{i=1}^{m} q(\boldsymbol{\zeta}_2^{(i)}, \boldsymbol{e}_+^{(i)}) \cdot \prod_{j=1}^{s} f\left(\sum_{k=1}^{s} R_{j,k}\tilde{\zeta}_{1,k}(\boldsymbol{e}_+^{(i)}) + \sum_{k=s+1}^{sn} R_{j,k}\epsilon_{T-s+k}^{(i)}\right)}{\frac{1}{m}\sum_{i=1}^{m} \prod_{j=1}^{s} f\left(\sum_{k=1}^{s} R_{j,k}\tilde{\zeta}_{1,k}(\boldsymbol{e}_+^{(i)} + \sum_{k=s+1}^{sn} R_{j,k}\epsilon_{T-s+k}^{(i)}\right)}, \tag{29}$$

where $(\boldsymbol{\zeta}_2^{(i)}, \boldsymbol{e}_+^{(i)}) = (\epsilon_{T+1}^{(i)}, ..., \epsilon_{T+M}^{(i)}) = \boldsymbol{\epsilon}_+^{(i)}$, $i = 1, ..., m$, are independent draws from a distribution with density $f \times \cdots \times f$ (M copies). Forecasts can be obtained by modifying the two steps in the forecasting procedure of the previous section as follows

Step 1. Generate $(\boldsymbol{\zeta}_2^{(i)}, \boldsymbol{e}_+^{(i)}) = (\epsilon_{T+1}^{(i)}, ..., \epsilon_{T+M}^{(i)}) = \boldsymbol{\epsilon}_+^{(i)}$, $i = 1, ..., m$, as described below (29).

Step 2. For $h = 1, 2, ...$, apply (29) recursively with $q(\boldsymbol{\zeta}_2, \boldsymbol{e}_+)$ and $\hat{\mathsf{E}}_T\left(y_{T+h}\right)$ as defined in Step 2 of the previous section.

This simulation procedure is similar to that derived in the purely noncausal case in Section 3.1 to which it, in fact, reduces in that special case (for a detailed discussion of this issue, see Appendix A.3).

The simulation procedure described above can also be used to obtain interval forecasts and forecasts for the conditional distribution of components of y_{T+h} ($h \geq 1$). In the case of one-step-ahead forecasts the approximation derived for $\mathsf{E}_T\left(\mathbf{1}\left(y_{a,T+1} \leq x\right)\right)$ at the end of the preceding section still applies and implies that a forecast for the conditional cumulative distribution function of $y_{a,T+1}$ at point x can be obtained by choosing $q(\boldsymbol{\zeta}_2, \boldsymbol{e}_+)$ as in (28) and applying the preceding Steps 1 and 2. A forecast of the whole conditional cumulative distribution function and, furthermore, interval forecast for $y_{a,T+1}$ can be obtained as described at the end of Section 3.1. The general case of obtaining interval forecasts and forecasts for the conditional distribution of the components of y_{T+h} with $h \geq 1$ is

discussed in Appendix A.4.

4. Simulation study

4.1. Simulated processes

In this section, we examine the performance of our forecasting techniques by using Monte Carlo simulations and data generation processes (DGPs) based on bivariate models estimated for real data. The same data, comprised of quarterly U.S. inflation and the real marginal cost, is also used in the next section to provide an illustration of our forecasting techniques. As mentioned in the Introduction, inflation and the real marginal cost are variables extensively studied in the previous literature, although instead of the real marginal cost other variables have also been considered along with inflation (see, e.g., Gali and Gertler (1999), Canova (2007), Nason and Smith (2008), Gefang, Koop, and Potter (2012), and the references in therein).

Our quarterly data set, from the Federal Reserve Bank of St. Louis FRED databank, covers the period from 1955:1 to 2010:3. Inflation is computed as the log-difference of the seasonally adjusted GDP implicit price deflator and the real marginal cost is approximated by the real unit labor cost (for details, see Lanne and Luoto (2013)). We use the period from 1955:1 to 1989:4 to estimate VAR(r, s) models that will serve as DGPs in the subsequent Monte Carlo simulations. Throughout this paper, GAUSS 10 and its BHHH optimization routine in the CMLMT package are employed in estimation, simulation, and forecasting.

To specify a potentially noncausal VAR model we proceed along the lines discussed in Section 2.2 and first consider a Gaussian VAR(p) model. The conventional model selection criteria AIC and BIC as well as autocorrelation functions of the residuals suggested the order $p = 2$. However, the assumption of Gaussian errors could be rejected by the Q–Q plots of the residuals and, given uncorrelated residuals, by the clear autocorrelation in the squared residuals of the inflation equation. Thus, we consider second-order models, that is, VAR(r, s) models with $r + s = 2$ and, to capture the fat tails of the residual distribution, we choose the (bivariate)

t–distribution for the errors.

In the previous literature, the typical reaction to autocorrelation in squared residuals of a conventional causal VAR model has been to augment the model with GARCH errors. However, as theoretically demonstrated by Lanne and Saikkonen (2013, Sec. 2.3), one can expect to find autocorrelation in squared residuals when a causal VAR model is fitted to data generated by a (non-Gaussian and) noncausal VAR process (the same applies to a noncausal VAR model whose orders are misspecified). Thus, a potentially viable alternative to augmenting a causal VAR model with GARCH errors is to consider a noncausal VAR model.

Of the second-order models the VAR$(0, 2)$ model maximizes the likelihood function but only marginally compared to the VAR$(1, 1)$ model, whereas in terms of residual diagnostics the VAR$(1, 1)$ model performs slightly better, as the residuals of the VAR$(0, 2)$ model appear conditionally heteroskedastic. In the VAR$(1, 1)$ model, the estimates of the parameters $\Pi_{1,12}$ and $\Phi_{1,12}$ appear small compared to their standard errors and the same applies to the estimates of the parameters $\Phi_{1,12}$ and $\Phi_{2,12}$ in the VAR$(0, 2)$ model (we use $\Phi_{k,ij}$ to signify the (i, j) element of the matrix Φ_k with a similar notation used for Π_k). Restricting these parameters to zero also seems reasonable according to the likelihood ratio test (p–values 0.271 and 0.083 in the VAR$(1, 1)$ and VAR$(0, 2)$ models, respectively) and, in the case of the VAR$(0, 2)$ model, their imposition considerably improves the rather imprecise estimation of the degrees-of-freedom parameter of the t–distribution. The restrictions has no marked effect on the residual diagnostics of the two models but, interestingly, the maximum value of the likelihood function of the restricted VAR$(1, 1)$ model turns out to be slightly greater than that of the VAR$(0, 2)$ model. All in all, both of these restricted models perform reasonably well and will be used as DGPs in our simulation experiments and in the forecasting exercise of Section 5. The estimation results are presented in Table 1. Below, we shall also consider the conventional causal VAR(2) model for comparison and, to see how our forecasting procedures work in a higher order case, a fourth-order model will be briefly discussed.

Table 1: Estimation results of the VAR(0,2) and VAR(1,1) models for the U.S. inflation and real marginal cost.

Panel A: VAR(0,2) model								
	0.618	0		0.271	0		1.260	0.152
Φ_1	(0.094)	(-)	Φ_2	(0.090)	(-)	Σ	(0.209)	(0.091)
	0.064	0.999		-0.142	-0.065		0.152	0.609
	(0.063)	(0.088)		(0.061)	(0.086)		(0.091)	(0.101)
λ	5.801		logL	-371.741				
	(1.743)							
Panel B: VAR(1,1) model								
	-0.347	0		0.915	0		1.178	0.581
Π_1	(0.088)	(-)	Φ_1	(0.032)	(-)	Σ	(0.202)	(0.311)
	-0.257	0.929		0.562	0.041		0.581	0.868
	(0.119)	(0.033)		(0.253)	(0.089)		(0.311)	(0.317)
λ	5.305		logL	-371.222				
	(1.619)							

Notes: The numbers in the parentheses are standard errors based on the Hessian of the log-likelihood function. In the table, λ is the degrees-of-freedom parameter of the multivariate t-distribution and logL is the value of the maximized log-likelihood function.

It may be worth noting that the restrictions employed in the noncausal models in Table 1 are imposed on purely statistical grounds. As they imply that neither leads nor lags of the marginal cost (y_{2t}) appear in the equation of inflation (y_{1t}), one might think that the marginal cost has no effect on inflation forecasts. However, one should be cautious about making such an interpretation. To see the reason for this, consider the VAR$(0, 2)$ model whose moving average representation is such that y_{1t} (inflation) depends on $\epsilon_{1,t+j}$, whereas y_{2t} (marginal cost) depends on both $\epsilon_{1,t+j}$ and $\epsilon_{2,t+j}$ $(j \geq 0)$. Thus, as $\epsilon_{1,t+j}$ affects both inflation and the marginal cost, one cannot rule out the possibility that the marginal cost can help forecast $\epsilon_{1,t+j}$ and thereby inflation (see the (approximate) forecasting formula (12)). A similar argument applies to the VAR$(1, 1)$ model.

4.2. Simulation set-up

We simulate 10 000 realizations of length $T + 8$ from DGPs corresponding to the two estimated models in Table 1 (100 observations are discarded from the beginning and end of the simulated series to eliminate the impact of initialization effects). We estimate a causal VAR(2) model as well as the correct noncausal VAR$(1, 1)$ or VAR$(0, 2)$ model from the first T observations in each realization. Note that the estimated models are unrestricted, i.e., the restrictions $\Phi_{1,12} = \Phi_{2,12} = 0$ and $\Pi_{1,12} = \Phi_{1,12} = 0$ discussed above are not taken into account. The sample size T is set to 300, and the number of simulated realizations m employed in the noncausal forecasting procedures ranges from $m = 10\ 000$ to $m = 500\ 000$. Results of some robustness checks with the sample size $T = 100$ will also be reported. Based on the findings of Lanne et al. (2012b), the value of the truncation parameter M is set at 50 (essentially the same results are obtained with $M = 30$ and $M = 100$).

Point forecasts 1–8 periods ahead are constructed as described in Section 3. When the forecasts are based on the noncausal VAR$(1, 1)$ model and importance sampling is used we have to choose the auxiliary density function $\varphi(\boldsymbol{\zeta}_2)$. Following the discussion at the end of Section 3.2.1, our choice is the density function of $(\epsilon_T, \epsilon_{T+1})$ with the independent ϵ_T and ϵ_{T+1} having the bivariate t–distribution shown in Table 1 (Panel B). In the case of the forecasting procedure derived in Section 3.2.2 the assumed nonsingularity boils down to the nonsingularity of the matrix N_0 (see the beginning of Section 3.2.2 and note that now $n = 2$ and $s = 1$). Using the estimates in Table 1 and formulas in Appendix A.1 we find that the determinant of N_0 is 0.173, showing that the required nonsingularity holds.

4.3. Results

Table 2 presents the determinants of the mean-squared forecast error (MSFE) matrices (cf., e.g., Athanasopoulos and Vahid (2008)) obtained by simulating the VAR$(0, 2)$ and VAR$(1, 1)$ processes discussed in the preceding section with forecast horizons ranging from 1 to 8 periods. Results obtained for the MSFEs of the two individual forecasts are qualitatively very similar and available upon request.

Table 2: Determinants of the mean-squared forecast error matrices of the VAR(0,2) and VAR(1,1) models described in Table 1.

Horizon	1	2	3	4	5	6	7	8
m	VAR(0,2)							
10 000	1.774	4.974	9.724	14.397	19.816	24.650	29.779	33.178
100 000	1.751	4.927	9.643	14.276	19.700	24.512	29.576	32.884
200 000	1.751	4.936	9.623	14.269	19.712	24.487	29.592	32.959
500 000	1.749	4.921	9.626	14.273	19.681	24.453	29.555	32.860
m	VAR(1,1), importance sampling (Section 3.2.1)							
10 000	1.809	4.994	9.005	14.083	18.445	23.296	28.711	32.970
100 000	1.741	4.915	8.835	13.833	18.199	22.881	28.253	32.433
200 000	1.753	4.922	8.878	13.872	18.222	22.864	28.129	32.389
500 000	1.748	4.903	8.826	13.846	18.150	22.741	28.079	32.271
m	VAR(1,1), without importance sampling (Section 3.2.2)							
10 000	1.734	4.939	8.879	13.941	18.271	22.870	28.103	32.263
100 000	1.716	4.882	8.802	13.801	18.075	22.729	28.057	32.224
200 000	1.719	4.881	8.802	13.784	18.082	22.700	28.025	32.219
500 000	1.716	4.881	8.801	13.795	18.066	22.701	28.045	32.224

Notes: The entries are based on 10 000 replications. The sample size is T=300 and m is the number of simulated realizations (see Section 3). The truncation parameter M is set at 50 (see, e.g., (11)). In importance sampling, the auxiliary density function $\varphi(\boldsymbol{\zeta}_2)$ is chosen as discussed in Section 4.2. In the first panel, the DGP is the VAR(0,2) process while in the other two cases it is the VAR(1,1) process (see Table 1). The noncausal (VAR(0,2) and VAR(1,1)) models are estimated without taking the zero restrictions in the DGP into account.

Overall, the results show that there is a clear improvement in forecast accuracy when the number of simulated realizations m increases from 10 000 to 100 000 or 200 000. The improvement is much smaller when m increases from 200 000 up to 500 000. Whether importance sampling is used (Section 3.2.1) or not (Section 3.2.2) has only a minor effect on the results obtained for the VAR$(1,1)$ model. By and large, forecasts based on the correct assumption of the nonsingularity of the matrix N_0 are slightly more accurate. Altogether the results suggest that, in practice, m = 200 000 is a reasonable choice. This is much more than needed in

Table 3: Relative mean-squared forecast errors (MSFEs) of the VAR(1,1) and VAR(0,2) models described in Table 1 compared to the Gaussian causal VAR(2) model.

Model	Horizon							
	1	2	3	4	5	6	7	8
				MSFE, y_{1t}				
VAR(1,1), IS	0.987	0.990	0.992	0.986	0.985	0.989	0.989	0.988
VAR(1,1)	0.971	0.986	0.987	0.983	0.982	0.985	0.987	0.986
VAR(0,2)	0.964	0.985	0.987	0.991	0.997	0.997	0.995	0.999
				MSFE, y_{2t}				
VAR(1,1), IS	1.002	1.005	1.001	1.003	1.002	0.999	0.996	0.997
VAR(1,1)	0.999	1.000	0.999	1.000	0.999	0.997	0.996	0.995
VAR(0,2)	1.000	1.003	1.004	1.004	1.002	0.996	0.996	0.997
				Det				
VAR(1,1), IS	0.990	0.994	0.995	0.989	0.987	0.987	0.985	0.986
VAR(1,1)	0.970	0.986	0.986	0.983	0.979	0.980	0.982	0.981
VAR(0,2)	0.966	0.989	0.993	0.996	1.001	0.995	0.992	0.997

Notes: See the notes to Table 2. Entries below unity indicate the superiority of the noncausal models. IS refers to importance-sampling-based forecasts. The number of simulated realizations is m=200 000.

the univariate case where Lanne et al. (2012b) found the choice $m = 10\ 000$ to be sufficient.

Table 3 shows the determinants of MSFE matrices and the individual MSFEs of the (correct) VAR$(0, 2)$ and VAR$(1, 1)$ models relative to a (misspecified) causal VAR(2) model with Gaussian errors (using t–distributed errors instead of Gaussian errors yields very similar results). In the case of the VAR$(1, 1)$ model both the importance-sampling-based forecasts (indicated by IS) and those based on the (correct) assumption of the nonsingularity of the matrix N_0 are considered. The number of simulated realizations is $m = 200\ 000$. The relative determinants of the MSFE matrices are always below unity, implying that gains in the overall forecast accuracy of the two variables can be achieved by using the correct noncausal model instead of its causal representation. However, an inspection of the individual MSFEs indicates that the gains are mainly due to forecasting the first variable (y_{1t}),

whose relative MSFEs are below unity, whereas those of the second variable (y_{2t}) lie around unity ranging between 0.995 and 1.004.

As a robustness check of our forecasting procedures we also examined DGPs obtained by estimating the parameters of the two models in Table 1 without imposing the zero restrictions in the simulated DGPs. The results were very similar to those reported in Tables 2 and 3, so that removing zero restrictions from the DGPs had no essential effect on forecasting accuracy (detailed results are available upon request).

Next we discuss results obtained with the smaller sample size $T = 100$. Without showing detailed results we first note that choosing the number of simulated realizations as $m = 200\ 000$ was still found appropriate, and is used to obtain Table 4 which presents results similar to those in Table 3 for $T = 100$. The results of Table 4 show that the relative MSFEs between the noncausal models and the Gaussian VAR model are somewhat larger than reported in Table 3. This is most likely due to the fact that the use of the smaller sample size has increased the estimation uncertainty, thereby resulting in less accurate forecasts. Support for this perception is obtained by considering the MSFEs based on the true parameter values of the VAR$(0, 2)$ and VAR$(1, 1)$ models instead of their estimates. For $T = 100$ the use of the true parameter values gave, on average, about 10% smaller relative MSFEs for the individual forecasts than reported in Table 4. In the case of the determinant of the MSFE matrix the average differences were even close to 20%. For the larger sample size $T = 300$ these differences were only about 5% at maximum, implying that the effect of estimation uncertainty on forecast accuracy is considerably larger for the smaller sample size $T = 100$.

As a small illustration of the potential consequences of (incorrectly) using the forecasting procedure of Section 3.2.2 when the matrix N_0 is singular we consider the bivariate VAR$(1, 1)$ model with the coefficient matrices given at the beginning of Section 3.2. Table 5 reports the relative MSFEs between the two forecasting procedures with the number of simulated realizations $m = 200\ 000$ and with the sample sizes $T = 100$ and $T = 300$ (note that the simulation results are based on

Table 4: Relative mean-squared forecast errors (MSFEs) of the VAR(1,1) and VAR(0,2) models compared to the Gaussian causal VAR(2) model when T=100.

Model	Horizon							
	1	2	3	4	5	6	7	8
				MSFE, y_{1t}				
VAR(1,1), IS	1.014	1.014	1.005	1.002	0.996	0.994	0.982	0.989
VAR(1,1)	1.011	1.004	0.998	0.997	0.991	0.990	0.981	0.984
VAR(0,2)	0.974	1.008	1.007	1.008	1.008	1.001	1.006	1.004
				MSFE, y_{2t}				
VAR(1,1), IS	1.014	1.014	1.005	1.002	0.996	0.994	0.982	0.989
VAR(1,1)	1.016	1.022	1.018	1.015	1.008	1.004	1.000	1.003
VAR(0,2)	1.042	1.019	1.014	1.010	1.008	1.001	0.997	0.992
				Det				
VAR(1,1), IS	1.061	1.053	1.036	1.029	1.014	1.010	0.995	1.005
VAR(1,1)	1.024	1.024	1.014	1.011	0.998	0.994	0.982	0.987
VAR(0,2)	1.015	1.032	1.026	1.023	1.022	1.008	1.007	1.000

Notes: See the notes to Tables 2–3. The sample size is T=100 and the number of simulated realizations is m=200 000.

using estimates of N_0 which are nonsingular, as discussed in Section 3.2.2). The results show that the differences between the two procedures are minor (the figures range between 0.997 and 1.005 for $T = 100$ and 0.993 and 1.003 for $T = 300$). This admittedly very limited simulation experiment suggests that, at least in the case $r = s = 1$, falsely relying on the nonsingularity assumption and employing the forecasting procedure of Section 3.2.2 is not critical. More evidence on this matter is needed, however, before any far-reaching conclusions can be drawn.

We also examined a fourth-order model to see how the two forecasting procedures derived in Section 3.2 perform in a higher-order case. The results are reported in Table 6 (due to heavier computations the number of replications is 5000 in these simulations). The DGPs were again estimated from the same data (AIC suggested order four for causal models with t–distributed errors). Of the fourth-order models, a VAR$(1, 3)$ model maximized the likelihood function. However, according to estimation results, this model appeared overparameterized and did not perform well

Table 5: Relative mean-squared forecast errors (MSFEs) of the VAR(1,1) model obtained with importance sampling and incorrectly assuming the nonsingularity of the matrix N_0.

	Horizon							
	1	2	3	4	5	6	7	8
				T=100				
MSFE, y_{1t}	1.001	1.000	1.001	1.002	1.001	1.001	1.001	1.001
MSFE, y_{2t}	1.004	0.998	0.997	0.999	1.001	1.000	1.001	1.002
Det	1.005	0.998	0.999	0.999	1.001	1.001	1.003	1.003
				T=300				
MSFE, y_{1t}	0.996	0.993	0.995	0.996	0.996	0.995	0.997	0.999
MSFE, y_{2t}	1.003	1.003	0.998	0.996	0.997	0.997	0.995	0.997
Det	0.999	0.998	0.995	0.994	0.995	0.994	0.995	0.996

Notes: The values of the autoregressive coefficients are given at the beginning of Section 3.2. The error term has a t-distribution with covariance matrix $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ and the value of the degree-of-freedom parameter 5.00. The entries above unity indicate larger MSFEs for importance-sampling-based forecasts. The number of simulated realizations is m=200 000 and the sample size is T=100 or T=300.

in terms of residual diagnostics. As the parameters $\Pi_{1,12}$ and $\Phi_{j,12}$, $j = 1, 2, 3$, were rather imprecisely estimated we restricted them to zero. These restrictions correspond to those used in the VAR$(0, 2)$ and VAR$(1, 1)$ models above, and when they were imposed a reasonable fit was obtained. Thus, we use this restricted VAR$(1, 3)$ model as a DGP in the higher-order case. The auxiliary density function $\varphi(\boldsymbol{\zeta}_2)$ needed in importance sampling was chosen as described at the end of Section 3.2.2 (in this case, four times the density function of the bivariate t–distributed error term ϵ_t). Qualitatively the simulation results in Table 6 are similar to those obtained for the VAR$(1, 1)$ model in Table 3. In particular, whether importance sampling is used or not has no substantial effect on the forecast accuracy, and compared to the causal VAR(4) model the forecasts are more accurate.

Table 6: Relative mean-squared forecast errors (MSFEs) of the VAR(1,3) model compared to the Gaussian causal VAR(4) model.

Model	Horizon							
	1	2	3	4	5	6	7	8
				MSFE, y_{1t}				
VAR(1,3), IS	1.015	1.013	0.990	0.988	0.998	1.011	1.007	1.003
VAR(1,3)	0.977	0.990	0.984	0.984	0.994	1.005	1.002	1.000
				MSFE, y_{2t}				
VAR(1,3), IS	0.963	0.966	0.979	0.975	0.983	0.988	0.990	0.990
VAR(1,3)	0.956	0.963	0.974	0.973	0.980	0.986	0.988	0.987
				Det				
VAR(1,3), IS	0.982	0.981	0.975	0.967	0.983	0.999	0.997	0.994
VAR(1,3)	0.940	0.956	0.962	0.961	0.977	0.992	0.990	0.987

Notes: See the notes to Tables 2–3. The sample size is T=300 and the number of simulated realizations is m=200 000. The results are based on 5 000 replications.

5. Empirical illustration

In this section, we consider out-of-sample forecasting with the bivariate causal and noncausal VAR models introduced in Table 1. An issue of special interest is whether U.S. inflation forecasts can also in the VAR framework be improved by allowing for noncausality, in accordance with the findings of Lanne et al. (2012a, 2012b) based on univariate AR models. Their results may reflect the fact that omitted factors predictable by lagged values of inflation are contained in the error term of a univariate AR model and the error term of the noncausal AR model is predictable unlike its causal counterpart. As the real marginal cost included in our bivariate model could be such an omitted factor, it is of interest to see how inflation forecasts behave when the real marginal cost is explicitly included in the model.

We compute forecasts by using an expansive window of observations such that the models are re-estimated at each date with the estimation period augmented by one observation. Following Lanne et al. (2012b), the starting point of the out-of-sample forecasting period is set to 1990:1 and the last forecasts are computed

for 2010:3, so that forecasts are computed for 83 quarters. Based on the simulation results of the previous section, the number of simulated realizations m used in forecasting with noncausal VAR(r, s) models ($s \geq 1$) is set at $m = 200\ 000$.

Based on the model selection results of the previous section we consider second-order models. Table 7 presents the individual MSFEs and determinants of the MSFE matrices for the causal VAR(2) models with Gaussian (VAR(2)-N) and t–distributed (VAR(2)-t) errors, and for the noncausal VAR$(1, 1)$ and VAR$(0, 2)$ models. Note that now the restrictions $\Phi_{1,12} = \Phi_{2,12} = 0$ and $\Pi_{1,12} = \Phi_{1,12} = 0$ are imposed on the VAR$(1, 1)$ and VAR$(0, 2)$ models, respectively. In the causal VAR(2) model no restrictions are employed, as in model selection reasonable restrictions were not found (this particularly applies to the restrictions $\Pi_{1,12} = \Pi_{2,12} = 0$).

First consider the inflation forecasts that we are mostly interested in. Table 7 shows that the VAR$(1, 1)$ model yields the smallest MSFEs except for the two-quarter horizon where it is slightly outperformed by the VAR$(0, 2)$ model. Moreover, irrespective of the forecast horizon, the VAR$(1, 1)$ model outperforms the two causal VAR(2) models of which the VAR(2)-N model performs better and it also performs quite well in comparison with the VAR$(1, 1)$ model when the forecast horizon is short. However, when the forecast horizon is four quarters or more the VAR$(1, 1)$ model is clearly superior. According to the test of Diebold and Mariano (1995) and West (1996) the differences in the forecast accuracy between the VAR$(1, 1)$ and Gaussian VAR(2) models for inflation are statistically significant in most cases, even at the 1% significance level. In line with the simulation results of the previous section, the differences between the two forecasting methods in the case of the VAR$(1, 1)$ model are negligible.

As far as forecasting the marginal cost is concerned, especially the Gaussian VAR(2) model performs slightly better than the noncausal models with the exception of one-quarter forecasts where the VAR$(0, 2)$ yields the smallest MSFEs. The differences are not statistically significant, however, and the determinants of the MSFE matrices reported in Table 7 show that the noncausal models produce the

Table 7: Mean-squared forecast errors (MSFEs) of the second-order causal and noncausal VAR(r, s) models for the U.S. inflation and marginal cost data.

Model	Forecast horizon (quarters)							
	1	2	3	4	5	6	7	8
				MSFE, inflation				
VAR(2)-N	1.073	1.426	1.694	2.075	2.769	3.379	3.969	4.387
VAR(2)-t	1.080	1.455	1.756	2.168	2.908	3.554	4.209	4.655
VAR(1,1), IS	1.068	1.373*	1.499***	1.777***	2.365***	2.800***	3.188***	3.436***
VAR(1,1)	1.066	1.371**	1.518**	1.789***	2.371***	2.817***	3.216***	3.464***
VAR(0,2)	1.077	1.368**	1.675	2.123	2.806	3.372	3.882	4.290
				MSFE, marginal cost				
VAR(2)-N	0.838	1.346	2.210	3.053	4.414	5.921	7.463	9.286
VAR(2)-t	0.849	1.351	2.234	3.106	4.518	6.103	7.744	9.698
VAR(1,1), IS	0.849	1.384	2.319	3.223	4.622	6.173	7.675	9.482
VAR(1,1)	0.844	1.383	2.316	3.218	4.631	6.170	7.686	9.491
VAR(0,2)	0.831	1.397	2.335	3.259	4.675	6.248	7.804	9.607
				Det				
VAR(2)-N	0.887	1.779	2.984	4.731	8.709	13.358	18.990	24.595
VAR(2)-t	0.904	1.817	3.088	4.916	9.113	13.960	20.026	25.856
VAR(1,1), IS	0.902	1.779	2.829	4.302	7.877	11.650	16.062	20.435
VAR(1,1)	0.896	1.771	2.868	4.290	7.869	11.682	16.190	20.542
VAR(0,2)	0.883	1.765	3.066	5.008	9.039	13.580	19.075	24.634

Notes: The entries are the MSFEs and determinants of the MSFE matrices of causal VAR(2) and noncausal VAR(1,1) and VAR(0,2) models. N and t denote Gaussian and t-distributed errors, respectively, and IS refers to importance-sampling-based forecasts. The number of simulated realizations is m=200 000. The stars *,** and *** signify statistically significant differences at 10%, 5% and 1% levels in the test of Diebold and Mariano (1995) and West (1996) used to test for equal forecast accuracy between the noncausal model (VAR(1,1) or VAR(0,2)) and the causal Gaussian VAR(2)-N model for inflation and marginal cost.

best overall forecasts. In particular, in terms of this criterion, the purely noncausal VAR$(0, 2)$ model yields the most accurate forecasts for one and two quarters ahead whereas the VAR$(1, 1)$ model is the best when the forecast horizon is longer.

As an illustration of obtaining interval forecasts we consider the one-step-ahead

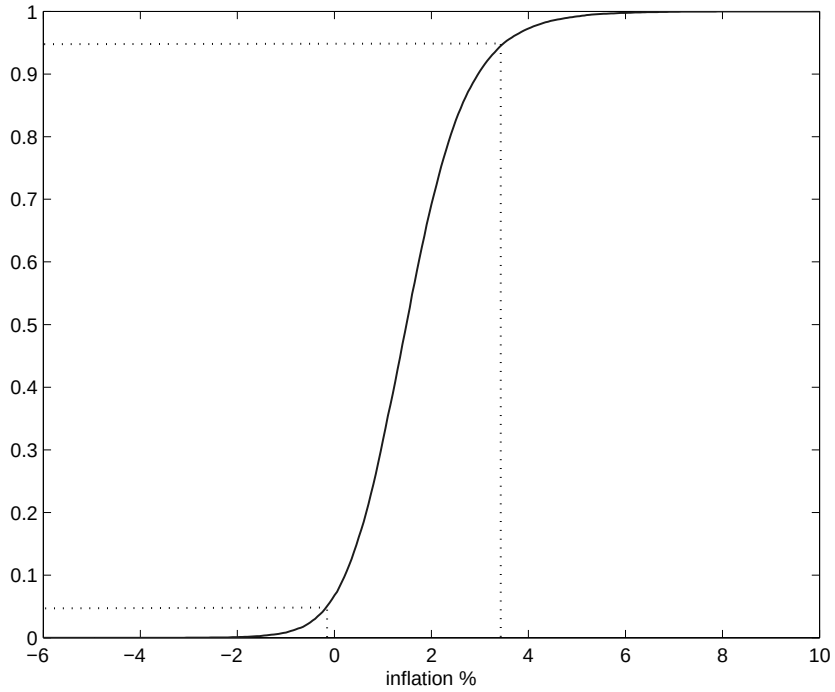

Figure 1: The conditional cumulative distribution function of inflation for the first quarter of the year 2008 predicted by the VAR(1,1) model. The dashed lines depict the lower and upper bounds of the 90% interval forecasts.

interval forecast of inflation. Figure 1 depicts a one-quarter-ahead forecast of the conditional cumulative distribution function of inflation for the first quarter of 2008. The forecast is formed by using the VAR$(1, 1)$ model with the importance-sampling-based forecasting technique (see Section 3.2.1). Any interval forecast can be read off the forecast of the conditional cumulative distribution function. The dotted lines show the lower (-0.15%) and upper (3.50%) bounds of the 90% interval forecasts when the point forecast is 1.87% and the realized value is 1.77%. Thus, the 90% forecast interval contains the observed inflation rate.

To sum up, the results show that the noncausal models produce more accurate forecasts for U.S. inflation than their causal alternatives, and this also holds for the bivariate system consisting of inflation and the marginal cost. However, causal models, especially the Gaussian model, perform slightly better than the noncausal

models in forecasting the marginal cost.

6. Conclusion

In this paper, we have developed forecasting methods for the noncausal VAR model of Lanne and Saikkonen (2013). To our knowledge, this is the first attempt to make forecasting in noncausal VAR models practically feasible. Due to the nonlinear nature of the prediction problem explicit formulas to compute forecasts are not available and, therefore, our forecasting methods exploit simulation-based techniques. The needed techniques turned out to be more complex than in the univariate case of Lanne et al. (2012b) with the extent of complexity depending on the structure of the model. However, according to the simulation experiments conducted, the proposed forecasting methods perform quite well even in the most complicated case, where importance sampling is employed. They also appear feasible in practice, as illustrated by our empirical application where noncausal VAR models performed well in comparison with their causal counterparts.

By making forecasting in the noncausal VAR model of Lanne and Saikkonen (2013) feasible in practice this paper has paved the way for developing methods for structural analysis within these models, including the computation of impulse response functions. Lanne and Saikkonen (2013) have also pointed out that noncausality is closely related to possible nonfundamental solutions of theoretical economic and financial models such as Dynamic Stochastic General Equilibrium (DSGE) models. As nonfundamentalness implies dependence on future error terms, it would be interesting to use the noncausal VAR model instead of the causal VAR model as a benchmark in assessing forecasting ability of DSGE models (cf. Rubaszek and Skrzypczynski, 2008).

Appendix: Technical details

A.1: Structure of the matrices N_j in (7)

In this appendix, we demonstrate that the matrix $[\boldsymbol{N}_0 \ \cdots \ \boldsymbol{N}_{sn-1}]$ $(sn \times sn^2)$ is of full row rank sn. First, conclude from the identity $\Phi(z^{-1})^{-1}\Xi(z) = N(z^{-1})$

that, as $\Xi(z) = I_n - \Xi_1 z - \cdots - \Xi_{(n-1)r} z^{(n-1)r}$,

$$\begin{aligned}
N_{-(n-1)r} &= -\Xi_{(n-1)r} \\
N_{-(n-1)r+1} &= \Phi_1 N_{-(n-1)r} - \Xi_{(n-1)r-1} \\
&\vdots \\
N_{-(n-1)r+s} &= \Phi_1 N_{-(n-1)r+s-1} + \cdots + \Phi_s N_{-(n-1)r} - \Xi_{(n-1)r-s} \\
&\vdots \\
N_{-1} &= \Phi_1 N_{-2} + \cdots + \Phi_s N_{-1-s} - \Xi_1.
\end{aligned}$$

Here, as well as elsewhere, $N_k = 0$ for $k < -(n-1)r$. Furthermore, the matrices N_k, $k \geq 0$, satisfy

$$\begin{aligned}
N_0 &= \Phi_1 N_{-1} + \cdots + \Phi_s N_{-s} + I_n \\
N_k &= \Phi_1 N_{k-1} + \cdots + \Phi_s N_{k-s}, \quad k \geq 1.
\end{aligned}$$

Note that in the pure noncausal case only the matrices N_j, $j \geq 0$, are relevant and the preceding equations apply with $N_j = 0$, $j < 0$. Because the matrices Ξ_j, $j = 1, ..., (n-1)r$, are functions of the parameters $\Pi_1, ..., \Pi_r$ the preceding equations show how the coefficient matrices N_j can be computed as functions of the autoregressive parameters.

Define the matrix

$$\boldsymbol{\Phi} = \begin{bmatrix}
\Phi_1 & \Phi_2 & \cdots & \Phi_{s-1} & \Phi_s \\
I_n & 0 & & & 0 \\
0 & \ddots & \ddots & & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & I_n & 0
\end{bmatrix} \quad (sn \times sn).$$

Then, using the definition of the matrix $\boldsymbol{N}_k$ (see the beginning of Section 3.2.1) we have

$$\boldsymbol{N}_k = \boldsymbol{\Phi}\boldsymbol{N}_{k-1} = \boldsymbol{\Phi}^k \boldsymbol{N}_0, \quad k \geq 1.$$

First we demonstrate that the rows of the infinite dimensional matrix $[\boldsymbol{N}_0 \ \boldsymbol{N}_1 \ \cdots]$ are linearly independent. As the spectral density matrix of y_t is positive definite

there can be no exact linear dependences between the components of the data vector $\boldsymbol{y}$. Thus, as the vector $\boldsymbol{z}$ is obtained from $\boldsymbol{y}$ by a nonsingular linear transformation (see the discussion preceding (10)) it follows that there can be no exact linear dependences between the components of $\boldsymbol{z}$. Hence, the same is true for $\boldsymbol{z}_3 = (v_{1,T-s+1}, ..., v_{s,T})$ and, as $v_{k,T-s+k} = \sum_{j=-k+1}^{\infty} N_j \epsilon_{T-s+k+j}$, we have

$$\begin{bmatrix} v_{1,T-s+1} \\ \vdots \\ v_{s,T} \end{bmatrix} = [\boldsymbol{N}_0 \ \cdots \ \boldsymbol{N}_{s-1} \ \boldsymbol{N}_s \ \cdots] \begin{bmatrix} \epsilon_{T-s+1} \\ \vdots \\ \epsilon_T \\ \epsilon_{T+1} \\ \vdots \end{bmatrix}.$$

From this it follows that the rows of the infinite dimensional matrix $[\boldsymbol{N}_0 \ \boldsymbol{N}_1 \ \cdots]$ are linearly independent.

Now we can proceed as in Hannan and Deistler (1988, p. 44-45). By the Caley-Hamilton theorem, the matrix $\boldsymbol{\Phi}$ satisfies its characteristic equation $\det(\mu I_{sn} - \boldsymbol{\Phi}) = 0$, which is of degree sn, so that $\boldsymbol{\Phi}^{sn} = c_1 I_{sn} + c_2 \boldsymbol{\Phi} + \cdots + c_{sn-1} \boldsymbol{\Phi}^{sn-1}$ for some scalars $c_1, ..., c_{sn-1}$. Thus, as $\boldsymbol{N}_k = \boldsymbol{\Phi}^k \boldsymbol{N}_0$, $k \geq 1$, we also have $\boldsymbol{N}_{sn} = c_1 \boldsymbol{N}_0 + c_2 \boldsymbol{N}_1 + \cdots + c_{sn-1} \boldsymbol{N}_{sn-1}$, implying that the columns of the matrix $\boldsymbol{N}_{sn}$ can be expressed as linear combinations the columns of the matrix $[\boldsymbol{N}_0 \ \cdots \ \boldsymbol{N}_{sn-1}]$. This fact can be extended inductively to the columns of any $\boldsymbol{N}_k$, $k \geq sn$. Thus, the matrix $[\boldsymbol{N}_0 \ \cdots \ \boldsymbol{N}_{sn-1}]$ must be of full row rank sn because otherwise we could find a vector c $(sn \times 1)$ such that $c' [\boldsymbol{N}_0 \ \boldsymbol{N}_1 \ \cdots] = 0$.

Note that the preceding discussion also shows that the matrix $\boldsymbol{N}_0$ must be nonzero because otherwise we would have $N_k = 0$ for all $k \geq 0$, implying that the matrix $[\boldsymbol{N}_0 \ \cdots \ \boldsymbol{N}_{sn-1}]$ is zero.

A.2: Joint density in (21)

In this appendix, we justify the approximate expression given for the conditional density function $p((\boldsymbol{\zeta}_2, \boldsymbol{e}_+) \mid \boldsymbol{y})$ in (21). Recall that $\boldsymbol{\zeta}_2 = \sum_{i=0}^{sn-1} \boldsymbol{K}_i \epsilon_{T-s+1+i}$ (see (19)) and $\boldsymbol{e}_+ = (\epsilon_{T-s+sn+1}, ..., \epsilon_{T+M})$. Now, conclude from the expression of the density function of $\boldsymbol{y}$ in equation (10) and the discussion preceding that equation

that the joint density function of $(\boldsymbol{y}, \boldsymbol{\zeta}_2, \boldsymbol{e}_+)$ is

$$p(\boldsymbol{y}, \boldsymbol{\zeta}_2, \boldsymbol{e}_+) = h_{\boldsymbol{z}_1}(\tilde{\boldsymbol{z}}_1) \cdot \left(\prod_{t=r+1}^{T-s} f(\tilde{\epsilon}_t) \right) \cdot h_{\boldsymbol{z}_3, \boldsymbol{\zeta}_2, \boldsymbol{e}_+}(\tilde{\boldsymbol{z}}_3, \boldsymbol{\zeta}_2, \boldsymbol{e}_+) \cdot |\det(\boldsymbol{H}_1)|,$$

where $\tilde{\epsilon}_t = \Pi(B)\Phi(B^{-1})y_t$ and $h_{\boldsymbol{z}_3, \boldsymbol{\zeta}_2, \boldsymbol{e}_+}(\tilde{\boldsymbol{z}}_3, \boldsymbol{\zeta}_2, \boldsymbol{e}_+)$ signifies the joint density function of $\boldsymbol{z}_3$, $\boldsymbol{\zeta}_2$, and $\boldsymbol{e}_+$ (note that here independence of $(\boldsymbol{z}_1, \boldsymbol{z}_2)$ and $(\boldsymbol{z}_3, \boldsymbol{\zeta}_2, \boldsymbol{e}_+)$ has also been used). Dividing both sides of the preceding equation by the density function of $\boldsymbol{y}$ (see (10)) shows that the conditional density function of $(\boldsymbol{\zeta}_2, \boldsymbol{e}_+)$ given $\boldsymbol{y}$ is

$$p((\boldsymbol{\zeta}_2, \boldsymbol{e}_+) \mid \boldsymbol{y}) = \frac{h_{\boldsymbol{z}_3, \boldsymbol{\zeta}_2, \boldsymbol{e}_+}(\tilde{\boldsymbol{z}}_3, \boldsymbol{\zeta}_2, \boldsymbol{e}_+)}{h_{\boldsymbol{z}_3}(\tilde{\boldsymbol{z}}_3)} = \frac{h_{\boldsymbol{z}_3, \boldsymbol{\zeta}_2, \boldsymbol{e}_+}(\tilde{\boldsymbol{z}}_3, \boldsymbol{\zeta}_2, \boldsymbol{e}_+)}{\int\int h_{\boldsymbol{z}_3, \boldsymbol{\zeta}_2, \boldsymbol{e}_+}(\tilde{\boldsymbol{z}}_3, \boldsymbol{\zeta}_2, \boldsymbol{e}_+) d\boldsymbol{\zeta}_2 d\boldsymbol{e}_+}.$$

Thus, we need to derive the joint density of $\boldsymbol{z}_3 = (v_{1,T-s+1}, ..., v_{s,T})$ and $(\boldsymbol{\zeta}_2, \boldsymbol{e}_+)$. It is shown below that this problem can be reduced to the derivation of $h_{\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2}(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2)$, the joint density function of $\boldsymbol{\zeta}_1$ and $\boldsymbol{\zeta}_2$. Specifically, we have

$$h_{\boldsymbol{z}_3, \boldsymbol{\zeta}_2, \boldsymbol{e}_+}(\tilde{\boldsymbol{z}}_3, \boldsymbol{\zeta}_2, \boldsymbol{e}_+) \approx h_{\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2}(\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+), \boldsymbol{\zeta}_2) \cdot \prod_{t=T-s+sn+1}^{T+M} f(\epsilon_t), \tag{30}$$

where $\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+) = \tilde{\boldsymbol{z}}_3 - \sum_{j=sn}^{M+s-1} \boldsymbol{N}_j \epsilon_{T-s+j+1}$ is as in (21). As $\boldsymbol{R} = [R_{j,k}] = \boldsymbol{Q}^{-1}$ $(sn^2 \times sn^2)$ is the matrix of the linear transformation $(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2) \to (\epsilon_{T-s+1}, .., \epsilon_{T-s+sn})$ (see (19)) it follows that the density function $h_{\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2}(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2)$ is as given in (22).

We still need to demonstrate the approximation (30). First recall that $\boldsymbol{\zeta}_1 = \boldsymbol{z}_3 - \sum_{j=sn}^{\infty} \boldsymbol{N}_j \epsilon_{T-s+j+1}$ (see the discussion following equation (19)). Thus, as $\boldsymbol{e}_+ = (\epsilon_{T-s+sn+1}, ..., \epsilon_{T+M})$, we get the approximate relation

$$\begin{bmatrix} \boldsymbol{\zeta}_1 \\ \boldsymbol{\zeta}_2 \\ \boldsymbol{e}_+ \end{bmatrix} \approx \boldsymbol{C} \begin{bmatrix} \boldsymbol{z}_3 \\ \boldsymbol{\zeta}_2 \\ \boldsymbol{e}_+ \end{bmatrix},$$

where

$$\boldsymbol{C} = \begin{bmatrix} I_{sn} & 0 & -\boldsymbol{N}_{sn} & -\boldsymbol{N}_{sn+1} & \cdots & -\boldsymbol{N}_{M+s-1} \\ 0 & I_{sn(n-1)} & 0 & 0 & \cdots & 0 \\ 0 & 0 & I_n & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & \cdots & 0 & I_n \end{bmatrix}.$$

The matrix $\boldsymbol{C}$ is clearly nonsingular with unit determinant. Thus, it follows that, to a close approximation, the joint density function of $\boldsymbol{z}_3$ and $(\boldsymbol{\zeta}_2, \boldsymbol{e}_+)$ is as given in (30) (note that here independence of $(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2)$ and $\boldsymbol{e}_+ = (\epsilon_{T-s+sn+1}, ..., \epsilon_{T+M})$ is also used).

A.3: Simulation procedure in Section 3.2.2 when $r = 0$

In this appendix, we demonstrate that in the purely noncausal case ($r = 0$) the forecasting technique derived in Section 3.2.2 reduces to that derived in Section 3.1. To simplify notation, we give details in the case $s = 1$ only.

When $s = 1$ one can readily check that (see the beginning of Section 3.2.2)

$$\boldsymbol{R} = \boldsymbol{Q}^{-1} = \begin{bmatrix} I_n & -N_1 & -N_2 & \cdots & -N_{n-1} \\ 0 & I_n & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & I_n \end{bmatrix}.$$

This implies that $\det(\boldsymbol{R}) = 1$ and, as now $R_{1,1} = I_n$ and $R_{1,k} = -N_k$ $(k = 2, ..., n)$, the density function $h_{\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2}(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2)$ employed in Section 3.2.2 takes the form

$$h_{\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2}(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2) = f\left(\zeta_{1,1} - \sum_{k=2}^{n} N_k \epsilon_{T-1+k}\right) \prod_{j=2}^{n} f\left(\epsilon_{T-1+j}\right).$$

Here we need to replace $\zeta_{1,1}$ by $\tilde{\zeta}_{1,1}(\boldsymbol{e}_+) = \tilde{v}_{1,T} - \sum_{j=n}^{M} N_j \epsilon_{T+j}$ with $\tilde{v}_{1,T} = y_T$ (see the definition of $\tilde{\zeta}_{1,k}(\boldsymbol{e}_+)$ below (21), (5), and (9)). Thus, consider the expression

$$f\left(\tilde{\zeta}_{1,1}(\boldsymbol{e}_+) - \sum_{k=2}^{n} N_k \epsilon_{T-1+k}\right) = f\left(y_T - \sum_{j=1}^{M} N_j \epsilon_{T+j}\right) = f\left(\tilde{\epsilon}_T\left(\boldsymbol{\epsilon}_+\right)\right),$$

where the latter equality is obtained by specializing the definition of $\tilde{\epsilon}_T(\boldsymbol{\epsilon}_+)$ to the case $s = 1$ (see the arguments leading to (15) in Section 3.1). As now $(\boldsymbol{\zeta}_2, \boldsymbol{e}_+) = \boldsymbol{\epsilon}_+$, the Monte Carlo approximation (29) in Section 3.2.2 can be expressed as

$$\hat{\mathsf{E}}_T\left(q\left(\boldsymbol{\zeta}_2, \boldsymbol{e}_+\right)\right) = \frac{\sum_{i=1}^{m} q(\boldsymbol{\epsilon}_+^{(i)}) f(\tilde{\epsilon}_T(\boldsymbol{\epsilon}_+^{(i)}))}{\sum_{i=1}^{m} f(\tilde{\epsilon}_T(\boldsymbol{\epsilon}_+^{(i)}))},$$

which with $q\left(\boldsymbol{\zeta}_2, \boldsymbol{e}_+\right) = \sum_{j=0}^{M-h} N_j \epsilon_{T+h+j}$ equals the expression obtained for $\hat{\mathsf{E}}_T\left(y_{T+h}\right)$ in Section 3.1 in the case $s = 1$. This shows the desired result.

When $s > 1$ the matrix $\boldsymbol{Q}$ is of the form

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ 0 & I_{sn(n-1)} \end{bmatrix},$$

where $\boldsymbol{Q}_{11}$ is upper triangular with unit diagonal elements. Computing the inverse of $\boldsymbol{Q}$ and using arguments similar to those above one can again show the desired result. Details are omitted.

A.4: Multiperiod interval forecasts for the general VAR(r, s) model

Using equations (5) and (7) we first consider the approximate relation

$$y_{T+h} \approx a_1 y_{T+h-1} + \cdots + a_{nr} y_{T+h-nr} + \sum_{j=-(n-1)r}^{M-h} N_j \epsilon_{T+h+j}, \quad h \geq 1,$$

and, following Lanne et al. (2012b), write it in companion form as

$$\boldsymbol{Y}_{T+h} \approx \boldsymbol{A}\boldsymbol{Y}_{T+h-1} + J \sum_{j=-(n-1)r}^{M} N_j \epsilon_{T+h+j},$$

where $\boldsymbol{Y}_{T+h} = (y_{T+h}, ..., y_{T+h-nr+1})$ $(n^2 r \times 1)$,

$$\boldsymbol{A} = \begin{bmatrix} a_1 I_n & a_2 I_n & \cdots & a_{nr-1} I_n & a_{nr} I_n \\ I_n & 0 & & & 0 \\ 0 & \ddots & \ddots & & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & I_n & 0 \end{bmatrix} \quad \left(n^2 r \times n^2 r\right) \quad \text{and} \quad J = \begin{bmatrix} I_n \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \left(n^2 r \times n\right).$$

Using repetitive substitution we can write the preceding approximation as

$$\begin{aligned} \boldsymbol{Y}_{T+h} &\approx \boldsymbol{A}^h \boldsymbol{Y}_T + \sum_{i=0}^{h-1} \boldsymbol{A}^i J \sum_{j=-(n-1)r}^{M} N_j \epsilon_{T+h-i+j} \\ &= \boldsymbol{A}^h \boldsymbol{Y}_T + \sum_{i=0}^{h-1} \boldsymbol{A}^i J \sum_{j=-(n-1)r}^{-s-h+i} N_j \epsilon_{T+h-i+j} \\ &\quad + \sum_{i=0}^{h-1} \boldsymbol{A}^i J \sum_{j=-s+1-h+i}^{-s+sn-h+i} N_j \epsilon_{T+h-i+j} + \sum_{i=0}^{h-1} \boldsymbol{A}^i J \sum_{j=-s+sn-h+i+1}^{M-h+i} N_j \epsilon_{T+h-i+j}, \end{aligned}$$

where the equality is based on the decomposition used in (17). Furthermore, the first two terms in the last expression are functions of the data, whereas the third one can be approximated as

$$\sum_{i=0}^{h-1} \boldsymbol{A}^i J \sum_{j=-s+1-h+i}^{-s+sn-h+i} N_j \epsilon_{T+h-i+j} \approx \sum_{i=0}^{h-1} \boldsymbol{A}^i J \boldsymbol{P}_{h-i} \begin{bmatrix} \boldsymbol{z}_3 - \sum_{j=sn}^{M+s-1} \boldsymbol{N}_j \epsilon_{T-s+j+1} \\ \boldsymbol{\zeta}_2 \end{bmatrix},$$

where $\boldsymbol{P}_{h-i} = [N_{-s+1-h+i} \cdots N_{-s+sn-h+i}]\, \boldsymbol{Q}^{-1}$ and, as before, $N_k = 0$, $k < -(n-1)\, r$. The argument used to obtain this approximation is the same as the one leading to (20). As $y_{T+h} = J' \boldsymbol{Y}_{T+h}$, the preceding discussion implies that

$$\begin{aligned} \mathsf{E}_T\left(y_{T+h}\right) \;\approx\; & J' \boldsymbol{A}^h \boldsymbol{Y}_T + \sum_{i=0}^{h-1} J' \boldsymbol{A}^i J \sum_{j=-(n-1)r}^{-s-h+i} N_j \tilde{\epsilon}_{T+h-i+j} \\ & + \mathsf{E}_T \left(\sum_{i=0}^{h-1} J' \boldsymbol{A}^i J \boldsymbol{P}_{h-i} \begin{bmatrix} \tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+) \\ \boldsymbol{\zeta}_2 \end{bmatrix} \right) \\ & + \mathsf{E}_T \left(\sum_{i=0}^{h-1} J' \boldsymbol{A}^i J \sum_{j=-s+sn-h+i+1}^{M-h+i} N_j \epsilon_{T+h-i+j} \right), \end{aligned}$$

where $\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+) = \tilde{\boldsymbol{z}}_3 - \sum_{j=sn}^{M+s-1} \boldsymbol{N}_j \epsilon_{T-s+j+1}$ as before.

Our forecast for the conditional cumulative distribution function of the ath component of y_{T+h} is obtained as

$$\mathsf{E}_T\left(\mathbf{1}\left(y_{a,T+h} \le x\right)\right) \approx \mathsf{E}_T \left(\mathbf{1} \left(\begin{array}{c} \sum_{i=0}^{h-1} \iota_a' J' \boldsymbol{A}^i J \boldsymbol{P}_{h-i} \begin{bmatrix} \tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+) \\ \boldsymbol{\zeta}_2 \end{bmatrix} \\ + \sum_{i=0}^{h-1} \iota_a' J' \boldsymbol{A}^i J \sum_{j=-s-h+sn+i+1}^{M-h+i} N_j \epsilon_{T+h-i+j} \le x - \iota_a' \tilde{\kappa}_{T,h} \end{array} \right) \right),$$

where

$$\tilde{\kappa}_{T,h} = J' \boldsymbol{A}^h \boldsymbol{Y}_T + \sum_{i=0}^{h-1} J' \boldsymbol{A}^i J \sum_{j=-(n-1)r}^{-s-h+i} N_j \tilde{\epsilon}_{T+h-i+j}$$

is a function of the data and, similarly to $\tilde{\boldsymbol{z}}_3$ in $\tilde{\boldsymbol{\zeta}}_1(\boldsymbol{e}_+)$, is treated as fixed. The conditional expectation on the right hand side of the preceding approximation is of the form $\mathsf{E}_T\left(q(\boldsymbol{\zeta}_2, \boldsymbol{e}_+)\right)$ with $q(\boldsymbol{\zeta}_2, \boldsymbol{e}_+)$ defined by the indicator function therein. Thus, we can use this choice of $q(\boldsymbol{\zeta}_2, \boldsymbol{e}_+)$ in (27) and the subsequent Steps 1 and 2 in Section 3.2.1 to obtain an approximate forecast for the conditional cumulative distribution function of $y_{a,T+h}$ ($h \ge 1$) at point x. A forecast of the whole

conditional cumulative distribution function and interval forecast for $y_{a,T+h}$ can be obtained as described at the end of the Section 3.1. Furthermore, when the approach of Section 3.2.2 is applicable the same choice of $q(\boldsymbol{\zeta}_2, \boldsymbol{e}_+)$ and Steps 1 and 2 of that section based on the simulation procedure (29) apply and can be used to obtain a forecast of the conditional cumulative distribution function and interval forecast for $y_{a,T+h}$ $(h \geq 1)$.

Acknowledgements

The authors thank the editor, an associate editor, two anonymous referees, and Markku Lanne for useful comments that led to considerable improvements of the first draft of this paper. The comments obtained at the 21st Symposium of the Society for Nonlinear Dynamics and Econometrics in Milan (March 2013) and the 33rd Annual International Symposium on Forecasting in Seoul (June 2013) have also been helpful when revising the paper. The financial support from the Academy of Finland, the OP-Pohjola Group Research Foundation, and the Yrjö Jahnsson Foundation is gratefully acknowledged. Part of this research was done during the first author's research visit to the Faculty of Economics of the University of Cambridge and during the second author's research visit to the Bank of Finland, whose hospitality is gratefully acknowledged.

References

Athanasopoulos, G., Vahid, F., 2008. VARMA versus VAR for macroeconomic forecasting. Journal of Business and Economic Statistics 26, 237–252.

Breidt, F.J., Davis, R.A., Lii, K.S., Rosenblatt, M., 1991. Maximum likelihood estimation for noncausal autoregressive processes. Journal of Multivariate Analysis 36, 175–198.

Breidt, F.J., Hsu, N.J., 2005. Best mean square prediction for moving averages. Statistica Sinica 15, 427–446.

Canova, F., 2007. G-7 inflation forecasts: Random walk, Phillips curve or what else? Macroeconomic Dynamics 11, 1–30.

Davis, R., Song, L., 2010. Noncausal vector AR processes with application to financial time series. Unpublished manuscript, Columbia University, New York.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. Journal of Business and Economic Statistics 13, 134–144.

Gali, J., Gertler, M., 1999. Inflation dynamics: A structural econometric approach. Journal of Monetary Economics 44, 195–222.

Gefang, D., Koop, G., Potter, S.M., 2012. The dynamics of UK and US inflation expectations. Computational Statistics and Data Analysis 56, 3120–3133.

Geweke, J., 1996. Monte Carlo simulation and numerical integration. In: Amman, H., Kendrick, D., Rust, J. (Eds.), Handbook of Computational Economics. North-Holland, Amsterdam, pp. 731–800.

Hannan, E.J., Deistler, M., 1988. The Statistical Theory of Linear Systems. Wiley, New York.

Lanne, M., Luoto, J., 2013. Autoregression-based estimation of the new Keynesian Phillips curve. Journal of Economic Dynamics and Control 37, 561–570.

Lanne, M., Saikkonen, P., 2011. Noncausal autoregressions for economic time series. Journal of Time Series Econometrics 3, article 2.

Lanne, M., Saikkonen, P., 2013. Noncausal vector autoregression. Econometric Theory 29, 447–481.

Lanne, M., Luoma, A., Luoto, J., 2012a. Bayesian model selection and forecasting in noncausal autoregressive models. Journal of Applied Econometrics 27, 812–830.

Lanne, M., Luoto, J., Saikkonen, P., 2012b. Optimal forecasting of noncausal autoregressive time series. International Journal of Forecasting 28, 623–631.

Lanne, M., Nyberg, H., Saarinen, E., 2012c. Does noncausality help in forecasting economic time series? Economics Bulletin 32, 2849–2859.

Lof, M., 2013. Noncausality and asset pricing. Studies in Nonlinear Dynamics and Econometrics 17, 211–220.

Nason, J.M., Smith, G.W., 2008. Identifying the new Keynesian Phillips curve. Journal of Applied Econometrics 23, 525–551.

Rosenblatt, M., 2000. Gaussian and Non-Gaussian Linear Time Series and Random Fields. Springer-Verlag, New York.

Rubaszek, M., Skrzypczynski, P., 2008. On the forecasting performance of a small-scale DSGE model. International Journal of Forecasting 24, 498–512.

West, K.D., 1996. Asymptotic inference about predictive ability. Econometrica 64, 1067–1084.